One "Ruler" for All Languages: Multi-Lingual Dialogue Evaluation with Adversarial Multi-Task Learning

Xiaowei Tong[1,2], Zhenxin Fu[1], Mingyue Shang[1], Dongyan Zhao[1,2], Rui Yan[1,2]*
[1]Institute of Computer Science and Technology, Peking University, China
[2]Beijing Institute of Big Data Research, China
{tongxiaowei, fuzhenxin, shangmy, zhaody, ruiyan}@pku.edu.cn

Abstract

Automatic evaluating the performance of Open-domain dialogue system is a challenging problem. Recent work in neural network-based metrics has shown promising opportunities for automatic dialogue evaluation. However, existing methods mainly focus on monolingual evaluation, in which the trained metric is not flexible enough to transfer across different languages. To address this issue, we propose an adversarial multi-task neural metric (ADVMT) for multi-lingual dialogue evaluation, with shared feature extraction across languages. We evaluate the proposed model in two different languages. Experiments show that the adversarial multi-task neural metric achieves a high correlation with human annotation, which yields better performance than monolingual ones and various existing metrics.

1 Introduction

The open-domain dialogue system is of growing interest in the field of Natural Language Processing (NLP). Its central goal is communicating with humans coherently and meaningfully; it also has wide industrial applications like XiaoIce[1] from Microsoft. Significant efforts have been made in recent years, to develop large-scale non-task-oriented dialogue system [Serban *et al.*, 2016; Li *et al.*, 2016; Tian *et al.*, 2017; Yao *et al.*, 2017; Song *et al.*, 2018]. These models adopt end-to-end neural network systems to predict the next dialogue utterance by the maximum likelihood estimation (MLE), given the previous dialogue turns.

Meanwhile, previous research has developed some successful automatic evaluation metrics. For example, BLEU [Papineni *et al.*, 2002] and METEOR [Banerjee and Lavie, 2005] are proposed for machine translation. ROUGE [Lin, 2004] is widely used in automatic summarization. However, when it comes to open-domain dialogue evaluation, these metrics are shown to correlate poorly with human judgments [Liu *et al.*, 2016a]. Researchers have to use those word-overlap metrics as there are few alternative efficient metrics [Li *et al.*, 2016; Yan *et al.*, 2016]. Some researchers rely on the manual annotation to evaluate their models, but it is costly and time-consuming. Therefore, having an accurate automatic dialogue evaluation model is in great need.

Very recently, some efforts have been made to develop a neural network-based metric for dialogue evaluation [Lowe *et al.*, 2017]. It learns to predict a score of a reply given its query (previous user-issued utterance) and a groundtruth reply. This method requires massive manual annotation. RUBER [Tao *et al.*, 2018] tries to address the cost of annotation through negative sampling and incorporating with referenced method.

However, the above methods only extract features from monolingual corpus, in which the trained metrics are not flexible enough to transfer across different language evaluation tasks simultaneously. Besides, these methods do not exploit a multi-lingual representation to enrich the features for automatic dialogue evaluation.

In this paper, we propose an adversarial multi-task learning for multi-lingual dialogue evaluation by integrating shared knowledge from multi-lingual corpora. Specifically, we regard each monolingual evaluation as a single task and propose a shared-private model under the framework of multi-task learning [Caruana, 1998; Ben-David *et al.*, 2003]. The multi-task learning structure contains two kinds of spaces: private and shared. The private feature spaces are used to extract the language-specific properties while the shared feature spaces capture the language-invariant properties across languages. Besides, motivated by the success of adversarial learning in domain adaption [Ganin *et al.*, 2016; Bousmalis *et al.*, 2016; Chen *et al.*, 2017], we incorporate adversarial strategy with shared spaces to enhance their ability to extract the common underlying features, and avoid the shared feature spaces being contaminated by noise.

The contributions of this paper could be summarized as follows:

- Multi-task learning is first introduced for automatic dialogue evaluation. It extracts not only language-specific features in private spaces but also language-invariant features in shared spaces across languages.
- An adversarial strategy is used to strengthen the ability to extract language-invariant features in shared spaces, in which a new objective function for multi-lingual dialogue evaluation is also proposed.

*Corresponding author: Rui Yan (ruiyan@pku.edu.cn)
[1]http://www.msxiaoice.com/

We evaluated adversarial multi-task neural metric (ADVMT) on both English and Chinese evaluation tasks. Experiments show that our proposed metric significantly outperforms existing automatic metrics in terms of the Pearson and Spearman correlation with human judgements, and has a boosted performance with the help of each monolingual evaluation task.

2 Related Work

2.1 Automatic Evaluation Metrics

From the machine learning perspective, automatic evaluation metrics can be divided into non-learnable and learnable approaches. Non-learnable metrics typically measure the quality of generated sentences by heuristics (manually defined equations), such as BLEU, ROUGE and Greedy Matching [Rus and Lintean, 2012]. As the valid reply in dialogue systems are of high diversity under a given context, these metrics are shown to correlate poorly with human judgments [Liu *et al.*, 2016a] for dialogue systems.

Compared with non-learnable metrics, learnable metrics can integrate linguistic features to enhance the correlation with human judgments through supervised learning. Lowe *et al.* [2017] develops a neural network-based metric for dialogue evaluation. RUBER [Tao *et al.*, 2018] addresses the cost of annotation through negative sampling and incorporating with referenced method. However, these metrics are trained in monolingual corpus, which are not flexible enough to transfer across different languages. Different from the above methods, our proposed metric extracts features from multi-lingual corpus and could be applied to different language evaluation tasks simultaneously.

2.2 Multi-task Learning with Neural Networks

The main concept of multi-task learning [Caruana, 1998] is to extract the common underlying features between related tasks and to improve the performance of each task with the help of private features and shared knowledge through parallel training. In recent years, researchers have incorporated it with recurrent neural networks (RNN) to address various NLP problems [Collobert and Weston, 2008; Hashimoto *et al.*, 2017].

Liu *et al.* [2016b] proposes a generic multi-task framework, in which different tasks can share information by an external memory and communicate by a reading/writing mechanism. Inspired by the success of multi-task learning, we regard each monolingual evaluation as a single task and propose a shared-private model under the framework of multi-task learning for multi-lingual dialogue evaluation.

2.3 Adversarial Neural Networks

Adversarial neural network [Goodfellow *et al.*, 2014] includes a neural generator G and a discriminator D, which is trained to classify real data versus generated data. Recently, the idea of adversarial networks is applied to various NLP tasks.

Chen *et al.* [2016] applies adversarial deep averaging network to transfer sentiment knowledge learned from labeled English data to low-resource languages where only unlabeled data exists. Chen *et al.* [2017] proposes an adversarial multi-criteria learning for Chinese word segmentation by integrating shared knowledge from multiple segmentation criteria. Liu *et al.* [2017] introduces an adversarial multi-task learning framework, alleviating the shared and private latent feature spaces from interfering with each other. Motivated by the success of adversarial networks, under the framework of multi-task learning, we incorporate adversarial strategy with shared spaces to enhance their ability to extract language-invariant features and propose a new objective function for multi-lingual dialogue evaluation.

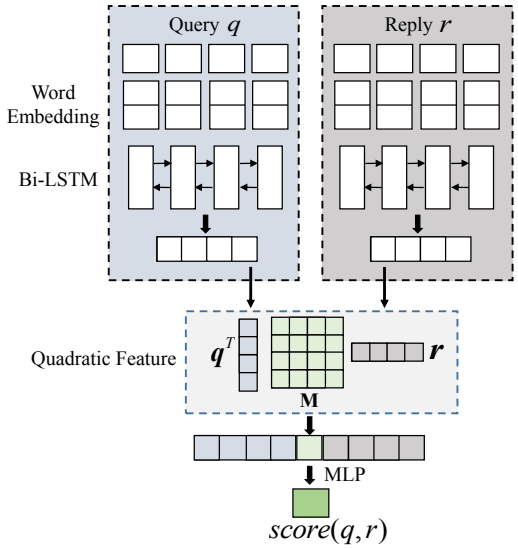

Figure 1: The single task neural metric to predict the score between a query q and its reply r.

3 Methodology

Given a previous query q and a reply r, the goal of neural network-based metric is to automatically measure the relatedness between q and r with a predicted $socre(q, r)$. In subsection 3.1 we introduce the neural network-based metric for monolingual dialogue evaluation, and regard it as a single task in our proposed multi-task learning framework in Subsection 3.2. In Subsection 3.3, we incorporate adversarial strategy to multi-task learning and introduce a new objective function for multi-lingual dialogue evaluation.

3.1 Neural Network-based Metric for Monolingual Dialogue Evaluation

This subsection mainly considers a single task neural metric to predict the appropriateness of a reply with respect to a query, of which the main structure (Figure 1) is inspired by Tao *et al.* [2018]. As for each word in a query q and a reply r, we first map them into vector representations (embedding). Then bi-directional Long Short-term Memory [Hochreiter and Schmidhuber, 1997] (Bi-LSTM) unites with forward and backward directions are applied to capture information along the word sequence. The update of each

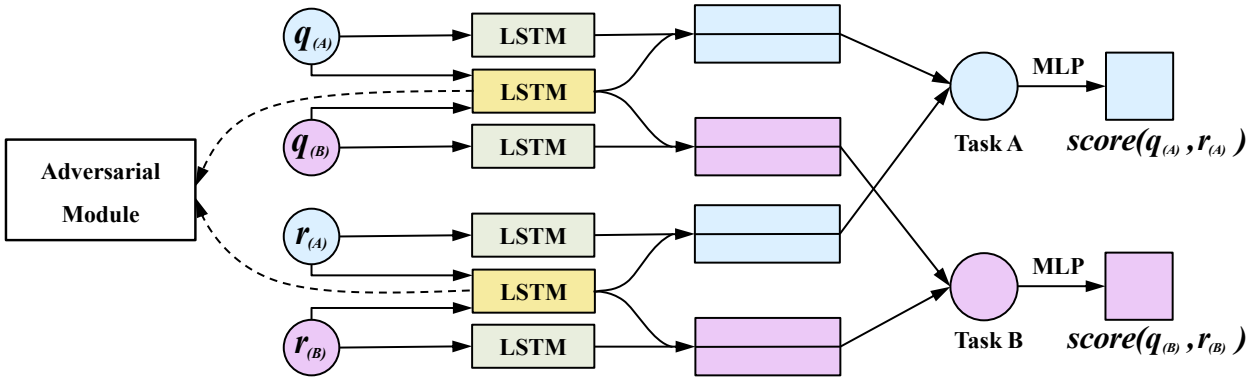

Figure 2: Overview of adversarial multi-task neural metric for multi-lingual dialogue evaluation. The blue and purple blocks indicate different language evaluation tasks A and B, respectively. The yellow LSTM blocks are shared spaces, while the gray LSTM blocks are private spaces.

Bi-LSTM unit can be written precisely as follows:

$$h_t = \overrightarrow{h}_t \oplus \overleftarrow{h}_t \quad (1)$$

$$= \text{Bi-LSTM}(x_t, \overrightarrow{h}_{t-1}, x_{T-t+1}, \overleftarrow{h}_{t-1}, \theta) \quad (2)$$

where $\boldsymbol{x_t}$ and T denote the embedding of the current input word and the last time step, and $\overrightarrow{h}_t$ is the forward hidden states. Likewise, $\overleftarrow{h}_t$ is the backward hidden states. $\oplus$ denotes the concatenation operation and all parameters in Bi-LSTM model is referred as θ. Specifically, we regard the concatenated output of both directions of Bi-LSTM, at the last time step T, as the representation of the whole sequence (q and r, respectively):

$$h_T^{(q)} = \text{Bi-LSTM}(x_T^{(q)}, \overrightarrow{h}_{T-1}^{(q)}, x_1^{(q)}, \overleftarrow{h}_{T-1}^{(q)}, \theta^{(q)}) \quad (3)$$

$$h_T^{(r)} = \text{Bi-LSTM}(x_T^{(r)}, \overrightarrow{h}_{T-1}^{(r)}, x_1^{(r)}, \overleftarrow{h}_{T-1}^{(r)}, \theta^{(r)}) \quad (4)$$

after which we concatenate $\boldsymbol{q}$ and $\boldsymbol{r}$ to match the two utterances. In addition, we include the "quadratic feature" that proposed in Tao *et al.* [2018], denoted as $\boldsymbol{q}^T\boldsymbol{M}\boldsymbol{r}$, where $\boldsymbol{M}$ is a parameter matrix. Finally we use a multi-layer perceptron (MLP) to predict a scalar score of the given conversation pairs. The MLP we adopted has two layers. The $tanh$ is used as the activation function in the hidden layers of MLP, while the second layer uses $sigmoid$ to make the score bounded.

In the process of training, we consider negative sampling to avoid costly manual annotation. Negative sampling is adopted for utterance matching in previous work [Yan *et al.*, 2016; 2017] and is shown to be feasible, which could ease the burden of costly manual annotation. Concretely, given a groundtruth query-reply pair, we randomly choose another reply r^- from the training set as a negative sample. The main goal of negative sampling training is to make the score of positive samples be larger than the negative samples by at least a margin δ. Thus the training objective is to minimize

$$J_{eval} = max\{0, \delta - score(q, r) + score(q, r^-)\} \quad (5)$$

3.2 Multi-task Learning for Multi-lingual Dialogue Evaluation

The method we introduced above only extract features from monolingual corpus, in which the trained metric is not flexible enough to transfer across different language evaluation tasks simultaneously and does not exploit a multi-lingual representation to enrich the features for automatic dialogue evaluation.

Inspired by the success of multi-task learning, we regard multi-lingual dialogue evaluation as multiple "related" tasks and propose a shared-private model, which shares information across languages. This shared-private mechanism is supposed to improve the performance of each other simultaneously with the help of shared features [Chen *et al.*, 2017; Liu *et al.*, 2017].

To enable multi-task learning for multi-lingual dialogue evaluation, as depicted in Figure 2, we design two feature spaces for both tasks A and B: a private space to capture language-dependent features, and a shared space to capture language-invariant features. Each monolingual evaluation task is assigned a private Bi-LSTM layer and a shared Bi-LSTM layer. Sentences are encoded by these two kinds of Bi-LSTM layers simultaneously.

Formally, for task k, the query vector representations of shared layer $q_T^{(s)}$ and private layer $q_T^{(p)}$ are formed as follows:

$$q_T^{(s)} = \text{Bi-LSTM}(x_T^{(s)}, \overrightarrow{q}_{T-1}^{(s)}, x_1^{(s)}, \overleftarrow{q}_{T-1}^{(s)}, \theta^{(s)}) \quad (6)$$

$$q_T^{(p)} = \text{Bi-LSTM}(x_T^{(p)}, \overrightarrow{q}_{T-1}^{(p)}, x_1^{(p)}, \overleftarrow{q}_{T-1}^{(p)}, \theta^{(p)}) \quad (7)$$

the reply vector representation of shared layer and private layer are denoted as $r_T^{(s)}$ and $r_T^{(p)}$, likewise. As all the tasks share the shared layer, the formula of shared layer is indicated by subscript.

To compute the similarity of query-reply pair, in each monolingual evaluation task, the sentence representations from private layer and shared layer are concatenated as the

final embedding. Specifically, for task k, the final sentence representations of query and reply are:

$$q_{(k)} = q_T^{(s)} \oplus q_T^{(p)} \quad (8)$$

$$r_{(k)} = r_T^{(s)} \oplus r_T^{(p)} \quad (9)$$

which are then concatenated to calculate the $score(q_{(k)}, r_{(k)})$ for each monolingual evaluation task k.

3.3 Incorporating Adversarial Strategy for Shared Spaces

Although the shared-private model separates the feature space into shared and private spaces, there is no guarantee that sharable features do not exist in private feature space, or vice versa [Liu *et al.*, 2017]. We hope that the features extracted by shared spaces is invariant across languages, under the multi-task learning framework for multi-lingual dialogue evaluation.

Inspired by the work on domain adaption [Ganin *et al.*, 2016; Bousmalis *et al.*, 2016], we exploit adversarial training strategy to optimize the shared layer, as shown in Figure 3. We use a discriminator to recognize which monolingual evaluation task the encoded sentence comes from. This discriminator maps the shared representation of sentences to a probability distribution, then makes a prediction of classes of monolingual evaluation tasks by its probability. The shared layers are designed to work defiantly towards a learnable multi-layer perceptron, preventing it from making an accurate prediction about the types of tasks. In this way, shared spaces are trained to be purer and less vulnerable to the contamination from private spaces.

Formally, for each monolingual evaluation task k, assume that there are N_k query-reply pairs. We refer to $s_{k,i}^{(q)}$ and $s_{k,i}^{(r)}$ as shared features from query and reply respectively, for i-th query-reply pair of task k.

We further concatenate $s_{k,i}^{(q)}$ and $s_{k,i}^{(r)}$ as the input of the discriminator, denoted as $s_{k,i} = s_{k,i}^{(q)} \oplus s_{k,i}^{(r)}$. Finally, the discriminator computes the probability distribution $P(k|s_{k,i}; \Theta_D, \Theta_S)$ as:

$$P(k|s_{k,i}; \Theta_D, \Theta_S) = softmax(W s_{k,i} + b) \quad (10)$$

where W is a learnable parameter and b is a bias; Θ_D are the parameters of discriminator; Θ_S indicate the parameters of shared spaces.

Based on such adversarial structure, besides the evaluation loss J_{eval}, we additionally introduce an adversarial loss, so that the discriminator could help to prevent shared spaces blending with task-specific features. The adversarial loss contains two parts: one is to train the discriminator to make an accurate prediction, and the other one aims to prevent the discriminator from predicting the class of monolingual evaluation tasks.

The task discriminator learns to determine which task the feature belongs to. Thus the training objective of it is to maximize the cross entropy of predicted task distribution. The loss function is formulated as follows:

$$J_{adv}^1(\Theta_D) = -\sum_{k=1}^{K}\sum_{i=1}^{N_k} logP(k|s_{k,i}; \Theta_D, \Theta_S) \quad (11)$$

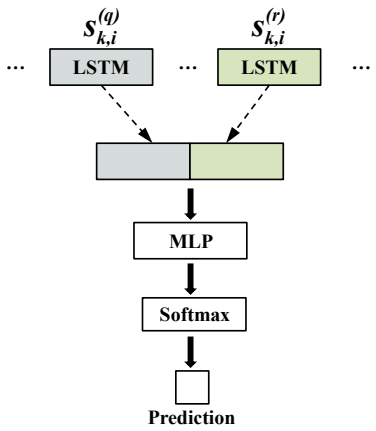

Figure 3: Architecture of adversarial training strategy for shared spaces. The blue and green LSTM blocks are shared layers from query $s_{k,i}^{(q)}$ and reply $s_{k,i}^{(r)}$ respectively, for i-th query-reply pair of task k.

where K denotes the evaluation tasks. It updates the parameters of discriminator Θ_D to minimize the loss function.

The other part of adversarial loss aims to prevent the discriminator from predicting the class of tasks. Therefore the training objective is:

$$J_{adv}^2(\Theta_S) = -\sum_{k=1}^{K}\sum_{i=1}^{N_k} -P(k|s_{k,i}) logP(k|s_{k,i}) \quad (12)$$

which is minimized by updating the parameters of shared layers Θ_S. $p(k|s_{k,i}; \Theta_D, \Theta_S)$ is referred as $p(k|s_{k,i})$ for short.

Combining the task evaluation loss and the adversarial loss, the final loss function of our adversarial multi-task neural metric for multi-lingual dialogue evaluation is defined as:

$$J = J_{eval} + J_{adv}^1 + J_{adv}^2 \quad (13)$$

where J_{eval} is computed in Eq (5).

4 Experiments

In this section, we evaluate the correlation between our proposed metrics and manual annotation, which is the ultimate goal of automatic metrics. Our model is trained with Chinese and English datasets under the adversarial multi-task neural network framework. The overall performance is investigated on Chinese and English corpus respectively.

4.1 Datasets

Chinese Corpus

We build a Chinese corpus using data crawled from an online Chinese forum Douban[2]. The training set contains 1,568,241 samples, each of which consists of a query-reply pair (in text). Standard Chinese word segmentation is applied to get Chinese terms as primitive tokens. We maintain a vocabulary of 129,506 phrases ranking by the term frequency, we empirically cut the phrases that frequency is under 3.

[2]http://www.douban.com/

Metrics		English Corpus (Twitter)		Chinese Corpus (Douban)	
		Pearson(p-value)	Spearman(p-value)	Pearson(p-value)	Spearman(p-value)
Inter-annotator	Human (Avg)	0.4478(<0.01)	0.4403(<0.01)	0.4692(<0.01)	0.4708(<0.01)
	Human (Max)	0.5510(<0.01)	0.5478(<0.01)	0.6068(<0.01)	0.6028(<0.01)
Referenced	BLEU-1	0.1214(<0.01)	0.0412(<0.01)	0.1521(<0.01)	0.2358(<0.01)
	BLEU-2	0.2016(<0.01)	0.2183(<0.01)	-0.0006(0.9914)	0.0546(0.3464)
	BLEU-3	0.1354(<0.01)	0.1701(<0.01)	-0.0576(0.3205)	-0.0188(0.7454)
	BLEU-4	0.2378(<0.01)	0.1324(<0.01)	-0.0604(0.2971)	-0.0539(0.3522)
	ROUGE	0.1702(<0.01)	0.0891(<0.01)	0.1747(<0.01)	0.2522(<0.01)
	Greedy Matching (GM)	0.2461(<0.01)	0.2388(<0.01)	0.3191(<0.01)	0.3137(<0.01)
Unreferenced	Single task	0.3685(<0.01)	0.3702(<0.01)	0.4071(<0.01)	0.4083(<0.01)
	Non-ADVMT	0.3823(<0.01)	0.3922(<0.01)	0.4249(<0.01)	0.4405(<0.01)
	ADVMT	**0.3901**(<0.01)	**0.4017**(<0.01)	**0.4317**(<0.01)	**0.4499**(<0.01)
RUBER	Min	0.3842(<0.01)	0.3721(<0.01)	**0.4527**(<0.01)	**0.4523**(<0.01)
	Geometric mean	**0.3928**(<0.01)	**0.3942**(<0.01)	0.4523(<0.01)	0.4490(<0.01)
	Arithmetic mean	0.3740(<0.01)	0.3688(<0.01)	0.4509(<0.01)	0.4458(<0.01)
	Max	0.3249(<0.01)	0.3126(<0.01)	0.3868(<0.01)	0.3623(<0.01)
ADVMT+GM	Min	0.4015(<0.01)	0.3981(<0.01)	0.4454(<0.01)	0.4535(<0.01)
	Geometric mean	**0.4267**(<0.01)	**0.4320**(<0.01)	**0.4698**(<0.01)	**0.4703**(<0.01)
	Arithmetic mean	0.3843(<0.01	0.3926(<0.01)	0.4170(<0.01)	0.4214(<0.01)
	Max	0.2908(<0.01)	0.3274(<0.01)	0.3991(<0.01)	0.3999(<0.01)

Table 1: Correlation between automatic metrics and human annotation. The p-value is a rough estimation of the probability that an uncorrelated metric produces a result that is at least as extreme as the current one; it does not indicate the degree of correlation.

English Corpora

We use the Twitter Corpus[3] that contains a large number of conversations between users on the micro-blogging platform Twitter as English Corpora. The training set contains 2,537,449 query-reply pairs. Like in Chinese corpus, we maintain a vocabulary of 125,291 words, of which the frequency is higher than 5.

4.2 Implementation Details

Hyperparameters

For sentence encoder, we set the word embedding size d_e for both tasks to 128, and they are initialized randomly and learned during training. The Bi-LSTM hidden states dimension d_h is set to 256 empirically. The learning rate α of Bi-LSTM units is initialized to 0.001. We use a two layers multi-layer perceptron to measure the relatedness of a given query and a reply. The dimension of the first layer d_m^1 is set to $8 * d_h + 1$ and dimension d_m^2 of the second layer is 50. The optimizer used in both Bi-LSTM and MLP are Adam [Kingma and Ba, 2014], and the gradient is computed by standard back-propagation. We set batch size (mini-batch) of both tasks to 128, and evaluate model on dev set after every 200 steps.

Performance Evaluation

We evaluate metrics on a generative model based on sequence-to-sequence (seq2seq) neural network [Bahdanau *et al.*, 2014]. This generative model encodes a query into a vector representation through a recurrent neural network (RNN), and decodes this vector into a reply with another RNN. To improve the performance of seq2seq model, attention mechanism is applied.

[3]http://www.twitter.com/

The English and Chinese test set include 300 queries and generated replies, respectively. We had 9 volunteers to express their human satisfaction of a generated reply to a query by rating an integer score among 0, 1 and 2. Score 2 means a "good" reply, 1 borderline, and 0 bad reply.

4.3 Results and Analysis

Table 1 shows the Pearson and Spearman correlation between some metrics and human scores. The evaluated metrics are as follows.

Referenced metrics predict the $score(r, \hat{r})$ between the ground-truth reply r and generated reply $\hat{r}$, including BLEU, ROUGE, and Greedy Matching (GM).

Unreferenced metrics include our Single Task, Non-adversarial Multi-task (Non-ADVMT) and Adversarial Multi-task (ADVMT) neural metrics. These metrics are unreferenced, because they predict the $score(q, \hat{r})$ between the query and its generated reply $\hat{r}$, without referring to a ground-truth reply r.

RUBER [Tao *et al.*, 2018] blends the referenced and unreferenced metrics by heuristics. For referenced $score(r, \hat{r})$ and unreferenced $score(q, \hat{r})$, it chooses the larger value (denoted as max), smaller value (min), and averaging (either geometric or arithmetic mean).

ADVMT+GM combines our Adversarial Multi-task neural metric and the Greedy Matching metric. The hybrid approach is the same as RUBER.

Overall Performance

The first observation in table 1 is that our unreferenced metrics are more correlated than those referenced metrics with human judgment in both English and Chinese evaluation tasks. This is because the referenced metrics mainly cap-

Query	Reply	Human score	Single-task	Non-ADVMT	ADVMT
He is not very popular.	OK!!!	0.19	0.30	0.24	0.20
So terrible today.	Ha-ha!	0.07	0.22	0.15	0.08
Where are you in Baoshan?	I'm in Minhang.	0.79	0.59	0.69	0.73
There is a teaching video.	Oh, thank you!	0.79	0.36	0.40	0.60
This is not an easy job.	Is there any value for this job?	0.36	0.11	0.17	0.32

Table 2: Selected cases with query and generated reply in Chinese and English, and Chinese is translated into English here. All the scores are mapped into the same section of $[0, 1]$ for directly comparing.

ture the similarity, but the rich semantic relationship between queries and replies necessitates more complicated mechanisms like neural networks. Besides, the referenced metrics have to rely on the information of both reference reply and generated reply, while neural network-based metrics use no reference but query and model reply. This observation shows that the query alone is also informative and that negative sampling could help to train the evaluation metrics, although it does not require human annotation as labels.

In unreferenced metrics, compared to single-task trained metric and non-adversarial multi-task trained metric, the ADVMT metric trained under the framework of adversarial multi-task achieves a best result in the correlation with human judgment. Although the corpora used in multi-task training are in different languages, the shared-private architecture shows the ability to extract useful language-invariant features. Thus when evaluating on a single language dialogue system, with the help of that shared information across languages, the performance is boosted. In addition, this result shows that incorporating the adversarial strategy could strengthen the ability to extract language-invariant features in shared spaces and help to prevent the shared spaces of features from being interfered by private spaces.

We combine the referenced metric GM and unreferenced metric ADVMT, and the hybrid approach is the same as RUBER. Experiments show that ADVMT+GM metric achieves the best result than RUBER peak performance, when choosing the geometric mean blended strategy. What's more, in both ADVMT+GM and RUBER metrics, choosing the larger value (max) is too lenient, and is slightly worse than other strategies. More importantly, our ADVMT metric is trained in multi-lingual dataset, which could be applied in multi-lingual dialogue evaluation simultaneously, while the RUBER metric should be trained in each monolingual dataset respectively and ignores massive information across languages.

Case Study

Table 2 illustrates some examples of Single Task, Non-ADVMT, ADVMT neural metric. As for the non-universal reply, we find that our ADVMT metric tends to give a closer score with the human score than the Single Task metric and Non-ADVMT metric. We further observe that a common problem of generative model is that it tends to generate a universal reply, such as "Ha-ha." and "Ok!" We observe that our ADVMT metric tends to give a lower score when it comes to such universal replies, while the Single Task metric and Non-ADVMT metric gives a relatively high score. Improving the diversity in generative model remains a challenging problem. But it may be furthered if the evaluation metrics used in training process encourage replies that of high diversities, and discourage those universal replies.

5 Conclusion and Future Work

In this paper, we propose an adversarial multi-task neural metric for multi-lingual dialogue evaluation, using shared feature extraction across languages. In addition, we incorporate adversarial strategy to shared spaces, which aims to guarantee the purity of shared feature spaces. Our proposed model regards models that trained in different language corpora as a single task and integrates each single task under the framework of adversarial multi-task learning. Experiments show that the proposed model outperforms the monolingual ones and various existing metrics.

An important direction of future research is evaluating the ability of the proposed metric to transfer knowledge from one language to another. There could be a problem of lacking training corpus when it comes to the dialogue system on minority languages. As the proposed metric could extract information across languages, the performance of multilingual evaluation metrics, which are trained on some majority languages with a massive corpus, of transferring the shared knowledge to the minority languages is worth exploring.

Acknowledgments

We appreciate the contribution from Chongyang Tao. Besides, we would like to thank the anonymous reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (No. 61672058). 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.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, volume 29, pages 65–72, 2005.

[Ben-David *et al.*, 2003] Shai Ben-David, Reba Schuller, et al. Exploiting task relatedness for multiple task learning. *Lecture notes in computer science*, pages 567–580, 2003.

[Bousmalis *et al.*, 2016] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351, 2016.

[Caruana, 1998] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

[Chen *et al.*, 2016] Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614*, 2016.

[Chen *et al.*, 2017] Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-criteria learning for chinese word segmentation. In *ACL (Volume 1: Long Papers)*, volume 1, pages 1193–1203, 2017.

[Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[Ganin *et al.*, 2016] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[Hashimoto *et al.*, 2017] Kazuma Hashimoto, Yoshimasa Tsuruoka, Richard Socher, et al. A joint many-task model: Growing a neural network for multiple nlp tasks. In *EMNLP*, pages 1923–1933, 2017.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *ACL*, pages 110–119, 2016.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL workshop*, volume 8. Barcelona, Spain, 2004.

[Liu *et al.*, 2016a] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, pages 2122–2132, 2016.

[Liu *et al.*, 2016b] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Deep multi-task learning with shared memory for text classification. In *EMNLP*, pages 118–127, 2016.

[Liu *et al.*, 2017] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Adversarial multi-task learning for text classification. In *ACL (Volume 1: Long Papers)*, volume 1, pages 1–10, 2017.

[Lowe *et al.*, 2017] Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. Towards an automatic turing test: Learning to evaluate dialogue responses. In *ACL (Volume 1: Long Papers)*, volume 1, pages 1116–1126, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. Association for Computational Linguistics, 2002.

[Rus and Lintean, 2012] Vasile Rus and Mihai Lintean. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 157–162. Association for Computational Linguistics, 2012.

[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.

[Song *et al.*, 2018] Yiping Song, Rui Yan, Yansong Feng, Yaoyun Zhang, Dongyan Zhao, and Ming Zhang. Towards a neural conversation model with diversity net using determinantal point processes. In *AAAI*, 2018.

[Tao *et al.*, 2018] Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI*, 2018.

[Tian *et al.*, 2017] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. How to make context more useful? an empirical study on context-aware neural conversational models. In *ACL (Volume 2: Short Papers)*, volume 2, pages 231–236, 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 *SIGIR*, pages 55–64. ACM, 2016.

[Yan *et al.*, 2017] Rui Yan, Dongyan Zhao, et al. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *SIGIR*, pages 685–694. ACM, 2017.

[Yao *et al.*, 2017] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP*, pages 2190–2199, 2017.