RUBER: An Unsupervised Method for Automatic Evaluation of Open-Domain Dialog Systems

Chongyang Tao,[1] **Lili Mou,**[2] **Dongyan Zhao,**[1] **Rui Yan**[1]
[1]Institute of Computer Science and Technology, Peking University, China
[2]Key Laboratory of High Confidence Software Technologies (Peking University)
Ministry of Education, China; Institute of Software, Peking University, China
{chongyangtao,zhaody,ruiyan}@pku.edu.cn
doublepower.mou@gmail.com

Abstract

Open-domain human-computer conversation has been attracting increasing attention over the past few years. However, there does not exist a standard automatic evaluation metric for open-domain dialog systems; researchers usually resort to human annotation for model evaluation, which is time- and labor-intensive. In this paper, we propose RUBER, a *Referenced metric and Unreferenced metric Blended Evaluation Routine*, which evaluates a reply by taking into consideration both a groundtruth reply and a query (previous user-issued utterance). Our metric is learnable, but its training does not require labels of human satisfaction. Hence, RUBER is flexible and extensible to different datasets and languages. Experiments on both retrieval and generative dialog systems show that RUBER has a high correlation with human annotation.

1 Introduction

Automatic evaluation is crucial to the research of open-domain human-computer dialog systems. Nowadays, open-domain conversation is attracting increasing attention as an established scientific problem (Bickmore and Picard, 2005; Bessho et al., 2012; Shang et al., 2015); it also has wide industrial applications like XiaoIce[1] from Microsoft and DuMi[2] from Baidu. Even in task-oriented dialog (e.g., hotel booking), an open-domain conversational system could be useful in handling unforeseen user utterances.

In existing studies, however, researchers typically resort to manual annotation to evaluate their models, which is expensive and time-consuming. Therefore, automatic evaluation metrics are particularly in need, so as to ease the burden of model comparison and to promote further research on this topic.

In early years, traditional vertical-domain dialog systems use metrics like slot-filling accuracy and goal-completion rate (Walker et al., 1997, 2001; Schatzmann et al., 2005). Unfortunately, such evaluation hardly applies to the open domain due to the diversity and uncertainty of utterances: "accurancy" and "completion," for example, make little sense in open-domain conversation.

Previous studies in several language generation tasks have developed successful automatic evaluation metrics, e.g., BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) for machine translation, and ROUGE (Lin, 2004) for summarization. For dialog systems, researchers occasionally adopt these metrics for evaluation (Ritter et al., 2011; Li et al., 2015). However, Liu et al. (2016) conduct extensive empirical experiments and show weak correlation between existing metrics and human annotation.

Very recently, Lowe et al. (2017) propose a neural network-based metric for dialog systems; it learns to predict a score of a reply given its query (previous user-issued utterance) and a groundtruth reply. But such approach requires massive human-annotated scores to train the network, and thus is less flexible and extensible.

In this paper, we propose RUBER, a *Referenced metric and Unreferenced metric Blended Evaluation Routine* for open-domain dialog systems. RUBER has the following distinct features:

- An embedding-based scorer measures the similarity between a generated reply and the groundtruth. We call this a *referenced* metric, because it uses the groundtruth as a reference, akin to existing evaluation metrics. Instead of

[1] http://www.msxiaoice.com/
[2] http://duer.baidu.com/

using word-overlapping information (e.g., in BLEU and ROUGE), we measure the similarity by pooling of word embeddings; it is more suited to dialog systems due to casual expressions in open-domain conversation.

- A neural network-based scorer measures the relatedness between the generated reply and its query. We observe that the query-reply relation is informative itself. This scorer is *unreferenced* because it does not refer to groundtruth. We apply negative sampling to train the network. Our approach requires no manual annotation label, and hence is more extensible than Lowe et al. (2017).
- We propose to combine the referenced and unreferenced metrics to better make use both worlds. On the one hand, closeness to groundtruth implies high quality. On the other hand, the groundtruth does not cover all possible ways of replying the query; the relatedness between a generated reply and its query then provides additional information. Combining these two aspects with heuristic strategies (e.g., averaging) further improves the performance.

In this way, RUBER does not require human annotation scores for training, in the sense of which, we call our metric *unsupervised*. Although we still have to prepare a corpus to train embeddings (in an unsupervised manner) and neural scorers (by negative sampling), the query-reply data—also a prerequisite in Lowe et al. (2017)—are much cheaper to obtain than human annotation of their satisfaction, showing the advantage of our approach.

We evaluated RUBER on prevailing dialog systems, including both retrieval and generative ones. Experiments show that RUBER significantly outperforms existing automatic metrics in terms of the Pearson and Spearman correlation with human judgments, and has fair transferability over different open-domain datasets.

2 Empirical Observations

In this section, we present our empirical observations regarding the question "*What makes a good reply in open-domain dialog systems?*"

Observation 1. Resembling the groundtruth generally implies a good reply. This is a widely adopted assumption in almost all metrics, e.g., BLEU, ROUGE, and METEOR. However, utterances are typically short and casual in dialog systems; thus word-overlapping statistics are of high variance. Candidate 1 in Table 1, for example, resembles the groundtruth in meaning, but shares only a few common words. Hence our method measures similarity based on embeddings.

Query:	为什么不领养一个?
	Why not adopt one?
Groundtruth:	我也想领养，但是不好找
	I hope so, but it's hard to find one
Candidate 1:	恩，我已经在朋友中找了
	I'm looking for one from my friends
Candidate 2:	巴比伦是稀有品种
	Pappilon is a rare breed
Candidate 3:	可以哈，谢谢你的建议
	OK, thank you for your advice

Table 1: Query and groundtruth/candidate replies.

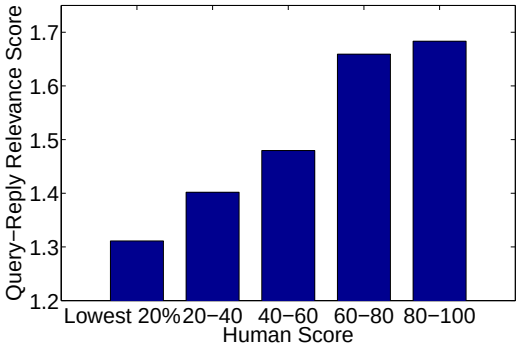


Figure 1: Average query-reply relevance scores versus quantiles of human scores. In other words, we divide human scores (averaged over all annotators) into 5 equal-sized groups, and show the average query-reply relevance score (introduced in Section 3.1) of each group.

Observation 2. A groundtruth reply is merely one way to respond. Candidate 2 in Table 1 illustrates a reply that is different from the groundtruth in meaning but still remains a good reply to the query. Moreover, a groundtruth reply may be universal itself (and thus undesirable). "I don't know,"—which appears frequently in the training set (Li et al., 2015)—may also fit the query, but it does not make much sense in a commercial chatbot.[3] The observation implies that a groundtruth alone is insufficient for the evaluation of open-domain dialog systems.

Observation 3. Fortunately, a query itself pro-

[3]Even if a system wants to mimic the tone of humans by saying "I don't know," it can be easily handled by post-processing. The evaluation then requires system-level information, which is beyond the scope of this paper.

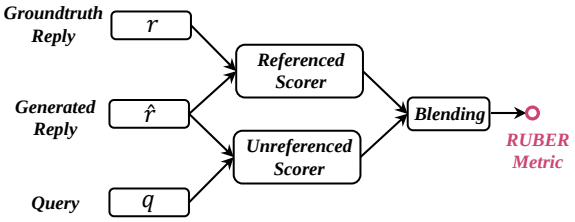


Figure 2: Overview of the RUBER metric.

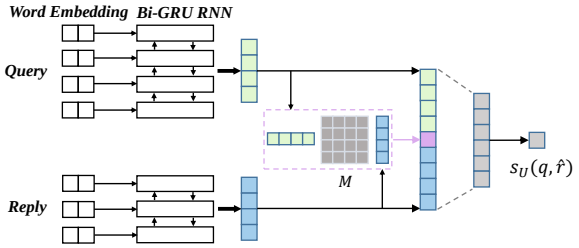


Figure 3: The neural network predicting the unreferenced score.

vides useful information in judging the quality of a reply.[4] Figure 1 plots the average human satisfactory score of a groundtruth reply versus the relevance measure (introduced in Section 3.2) between the reply and its query. We see that, even for groundtruth replies, those more relevant to the query achieve higher human scores. The observation provides rationales of using query-reply information as an unreferenced score in dialog systems.

3 Methodology

Based on the above observations, we design referenced and unreferenced metrics in Subsections 3.1 and 3.2, respectively; Subsection 3.3 discusses how they are combined. The overall design methodology of our RUBER metric is also shown in Figure 2.

3.1 Referenced Metric

We measure the similarity between a generated reply $\hat{r}$ and a groundtruth r as a referenced metric. Traditional referenced metrics typically use word-overlapping information including both precision (e.g., BLEU) and recall (e.g., ROUGE) (Liu et al., 2016). As said, they may not be appropriate for open-domain dialog systems.

We adopt the vector pooling approach that summarizes sentence information by choosing the maximum and minimum values in each dimension; the closeness of a sentence pair is measured by the cosine score. We use such heuristic matching because we assume no groundtruth scores, making it infeasible to train a parametric model.

Formally, let $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n$ be the embeddings of words in a sentence, max-pooling summarizes the maximum value as

$$\boldsymbol{v}_{\max}[i] = \max\left\{\boldsymbol{w}_1[i], \boldsymbol{w}_2[i], \cdots, \boldsymbol{w}_n[i]\right\} \quad (1)$$

where $[\cdot]$ indexes a dimension of a vector. Likewise, min pooling yields a vector $\boldsymbol{v}_{\min}$. Because an embedding feature is symmetric in terms of its sign, we concatenate both max- and min-pooling vectors as $\boldsymbol{v} = [\boldsymbol{v}_{\max}; \boldsymbol{v}_{\min}]$.

Let $\boldsymbol{v}_{\hat{r}}$ be the generated reply's sentence vector and $\boldsymbol{v}_r$ be that of the groundtruth reply, both obtained by max and min pooling. The referenced metric s_R measures the similarity between r and $\hat{r}$ by

$$s_R(r, \hat{r}) = \cos(\boldsymbol{v}_r, \boldsymbol{v}_{\hat{r}}) = \frac{\boldsymbol{v}_r^\top \boldsymbol{v}_{\hat{r}}}{\|\boldsymbol{v}_r\| \cdot \|\boldsymbol{v}_{\hat{r}}\|} \quad (2)$$

Forgues et al. (2014) propose a vector extrema method that utilizes embeddings by choosing either the largest positive or smallest negative value. Our heuristic here is more robust in terms of the sign of a feature.

3.2 Unreferenced Metric

We then measure the relatedness between the generated reply $\hat{r}$ and its query q. This metric is unreferenced and denoted as $s_U(q, \hat{r})$, because it does not refer to a groundtruth reply.

Different from the r-$\hat{r}$ metric, which mainly measures the similarity of two utterances, the q-$\hat{r}$ metric in this part involves more semantics. Hence, we empirically design a neural network (Figure 3) to predict the appropriateness of a reply with respect to a query.

Concretely, each word in a query q and a reply r is mapped to an embedding; a bidirectional recurrent neural network with gated recurrent units (Bi-GRU RNN) captures information along the word

[4] Technically speaking, a dialog generator is also aware of the query. However, a discriminative model (scoring a query-reply pair) is more easy to train than a generative model (synthesizing a reply based on a query). There could also be possibilities of generative adversarial training.

sequence. The forward RNN takes the form

$$
\begin{aligned}
[\boldsymbol{r}_t; \boldsymbol{z}_t] &= \sigma(W_{r,z}\boldsymbol{x}_t + U_{r,z}\boldsymbol{h}_{t-1}^{\rightarrow} + \boldsymbol{b}_{r,z}) \\
\tilde{\boldsymbol{h}}_t &= \tanh\left(W_h\boldsymbol{x}_t + U_h(\boldsymbol{r}_t \circ \boldsymbol{h}_{t-1}^{\rightarrow}) + \boldsymbol{b}_h\right) \\
\boldsymbol{h}_t^{\rightarrow} &= (1 - \boldsymbol{z}_t) \circ \boldsymbol{h}_{t-1}^{\rightarrow} + \boldsymbol{z}_t \circ \tilde{\boldsymbol{h}}_t
\end{aligned}
$$

where $\boldsymbol{x}_t$ is the embedding of the current input word, and $\boldsymbol{h}_t^{\rightarrow}$ is the hidden state. Likewise, the backward RNN gives hidden states $\boldsymbol{h}_t^{\leftarrow}$. The last states of both directions are concatenated as the sentence embedding ($\boldsymbol{q}$ for a query and $\boldsymbol{r}$ for a reply).

We further concatenate $\boldsymbol{q}$ and $\boldsymbol{r}$ to match the two utterances. Besides, we also include a "quadratic feature" as $\boldsymbol{q}^\top \mathbf{M} \boldsymbol{r}$, where $\mathbf{M}$ is a parameter matrix. Finally, a multi-layer perceptron (MLP) predicts a scalar score as our unreferenced metric s_U. The hidden layer of MLP uses $\tanh$ as the activation function, whereas the last (scalar) unit uses $\mathrm{sigmoid}$ because we hope the score is bounded.

The above empirical structure is mainly inspired by several previous studies (Severyn and Moschitti, 2015; Yan et al., 2016). We may also apply other variants for utterance matching (Wang and Jiang, 2016; Mou et al., 2016a); details are beyond the focus of this paper.

To train the neural network, we adopt negative sampling, which does not require human-labeled data. That is, given a groundtruth query-reply pair, we randomly choose another reply r^- in the training set as a negative sample. We would like the score of a positive sample to be larger than that of a negative sample by at least a margin Δ. The training objective is to minimize

$$
J = \max\left\{0, \Delta - s_U(q, r) + s_U(q, r^-)\right\} \quad (3)
$$

All parameters are trained by *Adam* (Kingma and Ba, 2014) with backpropagation.

In previous work, researchers adopt negative sampling for utterance matching (Yan et al., 2016). Our study further verifies that negative sampling is useful for the evaluation task, which eases the burden of human annotation compared with fully supervised approaches that requirer manual labels for training their metrics (Lowe et al., 2017).

3.3 Hybrid Approach

We combine the above two metrics by simple heuristics, resulting in a hybrid method RUBER for the evaluation of open-domain dialog systems.

We first normalize each metric to the range $(0, 1)$, so that they are generally of the same scale. In particular, the normalization is given by

$$
\tilde{s} = \frac{s - \min(s')}{\max(s') - \min(s')} \quad (4)
$$

where $\min(s')$ and $\max(s')$ refer to the maximum and minimum values, respectively, of a particular metric.

Then we combine $\tilde{s}_R$ and $\tilde{s}_U$ as our ultimate RUBER metric by heuristics including min, max, geometric averaging, and arithmetic averaging. As we shall see in Section 4.2, different strategies yield similar results, consistently outperforming baselines.

To sum up, RUBER metric is simple, general (without sophisticated model designs), and rather effective.

4 Experiments

In this section, we evaluate the correlation between our RUBER metric and human annotation, which is the ultimate goal of automatic metrics. The experiment was conducted on a Chinese corpus because of cultural background, as human aspects are deeply involved in this paper. We also verify the performance of RUBER metric when it is transferred to different datasets. We believe our evaluation routine could be applied to different languages.

4.1 Setup

We crawled massive data from an online Chinese forum Douban.[5] The training set contains 1,449,218 samples, each of which consists of a query-reply pair (in text). We performed Chinese word segmentation, and obtained Chinese terms as primitive tokens. In the referenced metric, we train 50-dimensional word2vec embeddings on the Douban dataset.

The RUBER metric (along with baselines) is evaluated on two prevailing dialog systems. One is a feature-based retrieval-and-reranking system, which first retrieves a coarse-grained candidate set by keyword matching and then reranks the candidates by human-engineered features; the top-ranked results are selected for evaluation (Song et al., 2016). The other is a sequence-to-sequence (Seq2Seq) neural network (Sutskever et al., 2014) that encodes a query as a vector with an RNN and

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

Metrics		Retrieval (Top-1)		Seq2Seq (w/ attention)	
		Pearson(p-value)	Spearman(p-value)	Pearson(p-value)	Spearman(p-value)
Inter-annotator	Human (Avg)	0.4927(<0.01)	0.4981(<0.01)	0.4692(<0.01)	0.4708(<0.01)
	Human (Max)	0.5931(<0.01)	0.5926(<0.01)	0.6068(<0.01)	0.6028(<0.01)
Referenced	BLEU-1	0.2722(<0.01)	0.2473(<0.01)	0.1521(<0.01)	0.2358(<0.01)
	BLEU-2	0.2243(<0.01)	0.2389(<0.01)	-0.0006(0.9914)	0.0546(0.3464)
	BLEU-3	0.2018(<0.01)	0.2247(<0.01)	-0.0576(0.3205)	-0.0188(0.7454)
	BLEU-4	0.1601(<0.01)	0.1719(<0.01)	-0.0604(0.2971)	-0.0539(0.3522)
	ROUGE	0.2840(<0.01)	0.2696(<0.01)	0.1747(<0.01)	0.2522(<0.01)
	Vector pool (s_R)	0.2844(<0.01)	0.3205(<0.01)	0.3434(<0.01)	0.3219(<0.01)
Unreferenced	Vector pool	0.2253(<0.01)	0.2790(<0.01)	0.3808(<0.01)	0.3584(<0.01)
	NN scorer (s_U)	0.4278(<0.01)	0.4338(<0.01)	0.4137(<0.01)	0.4240(<0.01)
RUBER	Min	0.4428(<0.01)	0.4490(<0.01)	**0.4527**(<0.01)	**0.4523**(<0.01)
	Geometric mean	0.4559(<0.01)	0.4771(<0.01)	0.4523(<0.01)	0.4490(<0.01)
	Arithmetic mean	**0.4594**(<0.01)	**0.4906**(<0.01)	0.4509(<0.01)	0.4458(<0.01)
	Max	0.3263(<0.01)	0.3551(<0.01)	0.3868(<0.01)	0.3623(<0.01)

Table 2: 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.

decodes the vector to a reply with another RNN; the attention mechanism (Bahdanau et al., 2015) is also applied to enhance query-reply interaction.

We had 9 volunteers to express their human satisfaction of a reply (either retrieved or generated) to a query by rating an integer score among 0, 1, and 2. A score of 2 indicates a "good" reply, 0 a bad reply, and 1 borderline.

4.2 Quantitative Analysis

Table 2 shows the Pearson and Spearman correlation between the proposed RUBER metric and human scores; also included are various baselines. Pearson and Spearman correlation are widely used in other research of automatic metrics such as machine translation (Stanojević et al., 2015). We compute both correlation based on q/r scores (either obtained or annotated), following Liu et al. (2016).

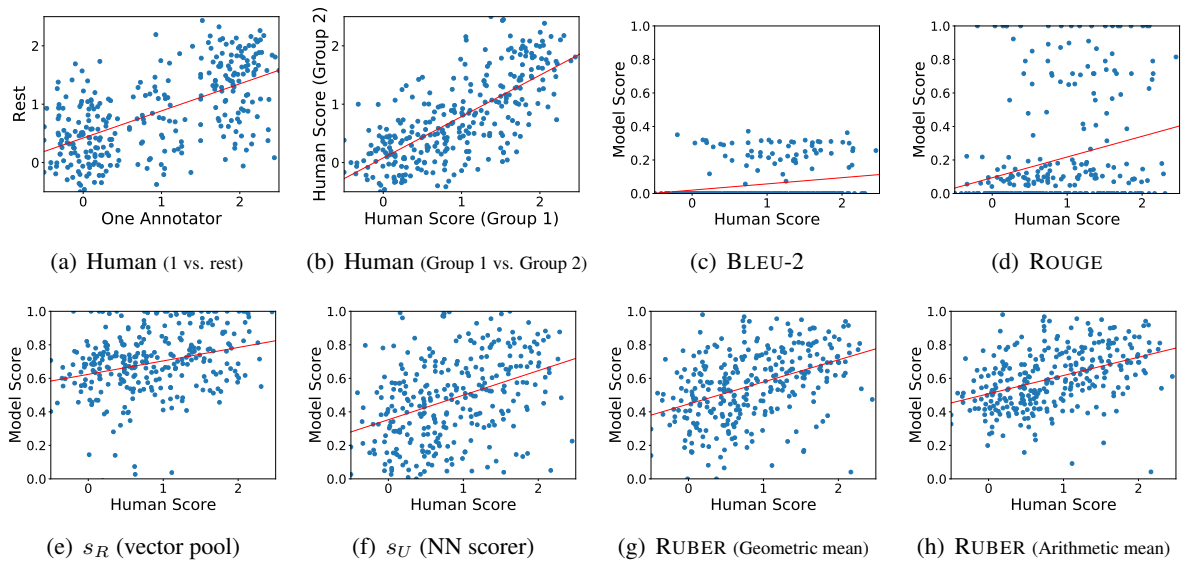
We find that the referenced metric s_R based on embeddings is more correlated with human annotation than existing metrics including both BLEU and ROUGE, which are based on word overlapping information. This implies the groundtruth alone is useful for evaluating a candidate reply. But exact word overlapping is too strict in the dialog setting; embedding-based methods measure sentence closeness in a "soft" way.

The unreferenced metric s_U achieves even higher correlation than s_R, showing that the query alone is also informative and that negative sampling is useful for training evaluation metrics, although it does not require human annotation as labels. Our neural network scorer outperforms the embedding-based cosine measure. This is because cosine mainly captures similarity, but the rich semantic relationship between queries and replies necessitates more complicated mechanisms like neural networks.

We combine the referenced and unreferenced metrics as the ultimate RUBER approach. Experiments show that choosing the larger value of s_R and s_U (denoted as $\max$) is too lenient, and is slightly worse than other strategies. Choosing the smaller value ($\min$) and averaging (either geometric or arithmetic mean) yield similar results. While the peak performance is not consistent in two experiments, they significantly outperforms both single metrics, showing the rationale of using a hybrid metric for open-domain dialog systems. We further notice that our RUBER metric has near-human correlation. More importantly, all components in RUBER are heuristic or unsupervised. Thus, RUBER does not requirer human labels; it is more flexible than the existing supervised metric (Lowe et al., 2017), and can be easily adapted to different datasets.

4.3 Qualitative Analysis

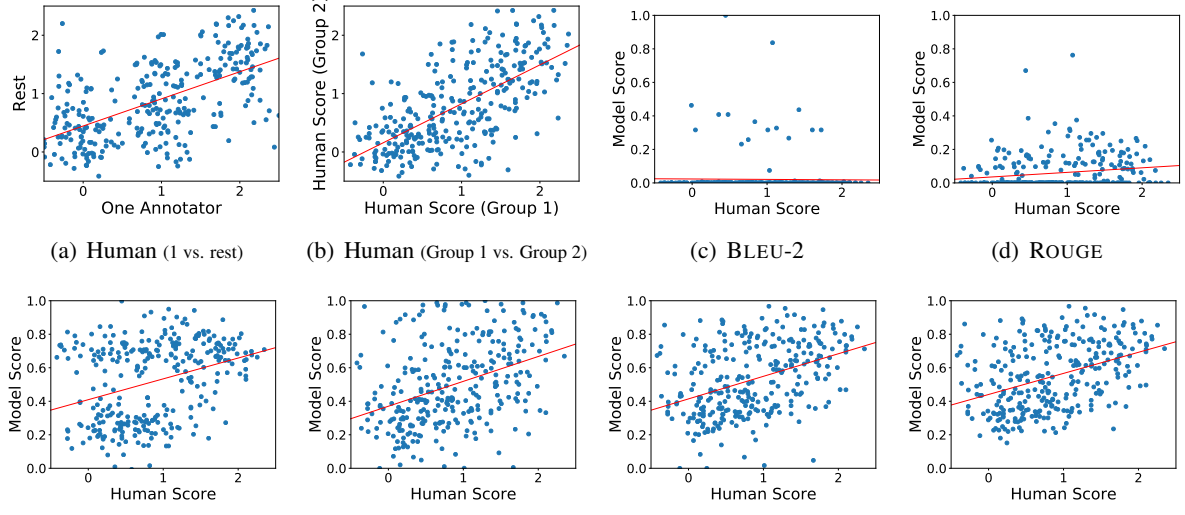
Figure 4 further illustrates the scatter plots against human judgments for the retrieval system, and Figure 5 for the generative system (Seq2Seq w/ attention). The two experiments yield similar results and show consistent evidence.



(a) Human (1 vs. rest) (b) Human (Group 1 vs. Group 2) (c) BLEU-2 (d) ROUGE

(e) s_R (vector pool) (f) s_U (NN scorer) (g) RUBER (Geometric mean) (h) RUBER (Arithmetic mean)

Figure 4: Score correlation of the retrieval dialog system. (a) Scatter plot of the medium-correlated human annotator against the rest annotators. (b) Human annotators are divided into two groups, one group vs. the other. (c)–(h) Scatter plots of different metrics against averaged human scores. Each point is associated with a query-reply pair; we add Guassian noise $\mathcal{N}(0, 0.25^2)$ to human scores for a better visualization of point density.



(a) Human (1 vs. rest) (b) Human (Group 1 vs. Group 2) (c) BLEU-2 (d) ROUGE

(e) s_R (vector pool) (f) s_U (NN scorer) (g) RUBER (Geometric mean) (h) RUBER (Arithmetic mean)

Figure 5: Score correlation of the generative dialog system (Seq2Seq w/ attention).

As seen, BLEU and ROUGE scores are zero for most replies, because for short-text conversation extract word overlapping occurs very occasionally; thus these metrics are too sparse. By contrast, both the referenced and unreferenced scores are not centered at a particular value, and hence are better metrics to use in open-domain dialog systems. Combining these two metrics results in a higher correlation (Subplots 4a and 5a).

We would like to clarify more regarding human-human plots. Liu et al. (2016) group human annotators into two groups and show scatter plots between the two groups, the results of which in our experiments are shown in Subplots 4b and 5b. However, in such plots, each data point's score is averaged over several annotators, resulting in low variance of the value. It is not a right statistic to compare with.[6] In our experimental de-

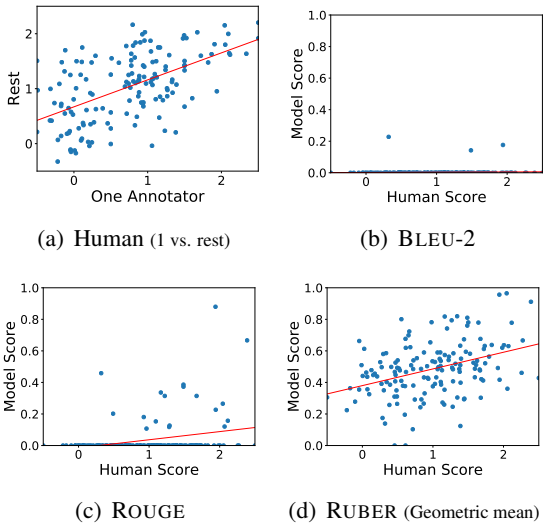
[6]In the limit of the annotator number to infinity, Subplots 4b and 5b would become diagonals (due to the Law of Large Numbers).

Query	Groundtruth Reply	Candidate Replies	Human Score	BLEU-2	ROUGE	s_U	s_R	RUBER
貌似离得挺近的 It seems very near.	你在哪里的嘞～ Where are you?	R1: 我也觉得很近 I also think it's near.	1.7778	0.0000	0.0000	1.8867	1.5290	1.7078
		R2: 你哪的？ Where are you from?	1.7778	0.0000	0.7722	1.1537	1.7769	1.4653

Table 3: Case study. In the third column, R1 and R2 are obtained by the generative and retrieval systems, resp. RUBER here uses arithmetic mean. For comparison, we normalize all scores to the range of human annotation, i.e., $[0, 2]$. Note that the normalization does not change the degree of correlation.

Metrics		Seq2Seq (w/ attention)	
		Pearson(p-value)	Spearman(p-value)
Inter-annotator	Human (Avg)	0.4860(<0.01)	0.4890(<0.01)
	Human (Max)	0.6500(<0.01)	0.6302(<0.01)
Referenced	BLEU-1	0.2091(0.0102)	0.2363(<0.01)
	BLEU-2	0.0369(0.6539)	0.0715(0.3849)
	BLEU-3	0.1327(0.1055)	0.1299(0.1132)
	BLEU-4	nan	nan
	ROUGE	0.2435(<0.01)	0.2404(<0.01)
	Vector pool (s_R)	0.2729(<0.01)	0.2487(<0.01)
Unreferenced	Vector pool	0.2690(<0.01)	0.2431(<0.01)
	NN scorer (s_U)	0.2911(<0.01)	0.2562(<0.01)
RUBER	Min	0.3629(<0.01)	0.3238(<0.01)
	Geometric mean	**0.3885**(<0.01)	**0.3462**(<0.01)
	Arithmetic mean	0.3593(<0.01)	0.3304(<0.01)
	Max	0.2702(<0.01)	0.2778(<0.01)

Table 4: Correlation between automatic metrics and human annotation in the transfer setting.



(a) Human (1 vs. rest) (b) BLEU-2

(c) ROUGE (d) RUBER (Geometric mean)

Figure 6: Score correlation of the generative dialog system (Seq2Seq w/ attention) in the transfer setting.

sign, we would like to show the difference between a single human annotator versus the rest annotators; in particular, the scatter plots 4a and 5a demonstrate the median-correlated human's performance. These qualitative results show our RUBER metric achieves similar correlation to humans.

4.4 Case Study

Table 3 illustrates an example of our metrics as well as baselines. We see that BLEU and ROUGE scores are prone to being zero. Even the second reply is very similar to the groundtruth, its Chinese utterances do not have bi-gram overlap, resulting in a BLEU-2 score of zero. By contrast, our referenced and unreferenced metrics are denser and more suited to open-domain dialog systems.

We further observe that the referenced metric s_R assigns a high score to R1 due to its correlation with the query, whereas the unreferenced metric s_U assigns a high score to R2 as it closely resembles the groundtruth. Both R1 and R2 are considered reasonable by most annotators, and our RUBER metric yields similar scores to human annotation by balancing s_U and s_R.

4.5 Transferability

We would like to see if the RUBER metric can be transferred to different datasets. Moreover, we hope RUBER can be directly adapted to other datasets even without re-training the parameters.

We crawled another Chinese dialog corpus from the Baidu Tieba[7] forum. The dataset comprises 480k query-reply pairs, and its topics may vary from the previously used Douban corpus. We only evaluated the results of the Seq2Seq model (with attention) because of the limit of time and space.

We directly applied the RUBER metric to the Baidu dataset, i.e., word embeddings and s_R's parameters were trained on the Douban dataset. We also had 9 volunteers to annotate 150 query-reply pairs, as described in Section 4.1. Table 4 shows the Pearson and Spearman correlation and Figure 6 demonstrates the scatter plots in the transfer setting.

As we see, transferring to different datasets leads to slight performance degradation compared with Table 2. This makes sense because the parameters, especially the s_R scorer's, are not trained for the Tieba dataset. That being said, RUBER still significantly outperforms baseline metrics, showing fair transferability of our proposed method.

[7] http://tieba.baidu.com

Regarding different blending methods, min and geometric/arithmetic mean are similar and better than the max operator; they also outperform their components s_R and s_U. The results are consistent with the non-transfer setting (Subsections 4.2 and 4.3), showing additional evidence of the effectiveness of our hybrid approach.

5 Related Work

5.1 Automatic Evaluation Metrics

Automatic evaluation is crucial to the research of language generation tasks such as dialog systems (Li et al., 2015), machine translation (Papineni et al., 2002), and text summarization (Lin, 2004). The Workshop on Machine Translation (WMT) organizes shared tasks for evaluation metrics (Stanojević et al., 2015; Bojar et al., 2016), attracting a large number of researchers and greatly promoting the development of translation models.

Most existing metrics evaluate generated sentences by word overlapping against a groundtruth sentence. For example, BLEU (Papineni et al., 2002) computes geometric mean of the precision for n-gram ($n = 1, \cdots, 4$); NIST (Doddington, 2002) replaces geometric mean with arithmatic mean. Summarization tasks prefer recall-oriented metrics like ROUGE (Lin, 2004). METEOR (Banerjee and Lavie, 2005) considers precision as well as recall for more comprehensive matching. Besides, several metrics explore the source information to evaluate the target without referring to the groundtruth. Popović et al. (2011) evaluate the translation quality by calculating the probability score based on IBM Model I between words in the source and target sentences. Louis and Nenkova (2013) use the distribution similarity between input and generated summaries to evaluate the quality of summary content.

From the machine learning perspective, automatic evaluation metrics can be divided into non-learnable and learnable approaches. Non-learnable metrics (e.g., BLEU and ROUGE) typically measure the quality of generated sentences by heuristics (manually defined equations), whereas learnable metrics are built on machine learning techniques. Specia et al. (2010) and Avramidis et al. (2011) train a classifier to judgment the quality with linguistic features extracted from the source sentence and its translation. Other studies regard machine translation evaluation as a regression task supervised by manually annotated scores (Albrecht and Hwa, 2007; Giménez and Márquez, 2008; Specia et al., 2009).

Compared with traditional heuristic evaluation metrics, learnable metrics can integrate linguistic features[8] to enhance the correlation with human judgments through supervised learning. However, handcrafted features often require expensive human labor, but do not generalize well. More importantly, these learnable metrics require massive human-annotated scores to learn the model parameters. Different from the above methods, our proposed metric apply negative sampling to train the neural network to measure the relatedness of query-reply pairs, and thus can extract features automatically without any supervision of human-annotated scores.

5.2 Evaluation for Dialog Systems

Dialog systems based on generative methods are also language generation tasks, and thus several researchers adopt BLEU score to measure the quality of a reply (Li et al., 2015; Sordoni et al., 2015; Song et al., 2016). However, its effectiveness has been questioned (Callison-Burch et al., 2006; Galley et al., 2015). Meanwhile, Liu et al. (2016) conduct extensive empirical experiments and show the weak correlation of existing metrics (e.g., BLEU, ROUGE and METEOR) with human judgements for dialog systems. Based on BLEU, Galley et al. (2015) propose ΔBLEU, which considers several reference replies. However, multiple references are hard to obtain in practice.

Recent advances in generative dialog systems have raised the problem of universally relevant replies. Li et al. (2015) measure the reply diversity by calculating the proportion of distinct unigrams and bigrams. Besides, Serban et al. (2016) and Mou et al. (2016b) use entropy to measure the information of generated replies, but such metric is independent of the query and groundtruth. Compared with the neural network-based metric proposed by Lowe et al. (2017), our approach does not require human-annotated scores.

6 Conclusion and Discussion

In this paper, we proposed an evaluation methodology for open-domain dialog systems. Our metric is called RUBER (a *Referenced metric and Unreferenced metric Blended Evaluation Routine*), as it

[8] Technically speaking, existing metrics (e.g., BLEU and METEOR) can be regarded as features extracted from the output sentence and the groundtruth.

considers both the groundtruth and its query. Experiments show that, although unsupervised, RUBER has strong correlation with human annotation, and has fair transferability over different open-domain datasets.

Our paper currently focuses on single-turn conversation as a starting point of our research. However, the RUBER framework can be extended naturally to more complicated scenarios: in a history/context-aware dialog system, for example, the modification shall lie in designing the neural network, which will take context into account, for the unreferenced metric.

References

Joshua Albrecht and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. pages 296–303.

Eleftherios Avramidis, Maja Popović, David Vilar, and Aljoscha Burchardt. 2011. Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. pages 65–70.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. pages 65–72.

Fumihiro Bessho, Tatsuya Harada, and Yasuo Kuniyoshi. 2012. Dialog system using real-time crowdsourcing and twitter large-scale corpus. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 227–231.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction* 12(2):293–327.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. pages 199–231.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluation the role of BLEU in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*. pages 249–256.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*. pages 138–145.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NIPS ML-NLP Workshop*.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and Bill Dolan. 2015. deltaBLEU: a discriminative metric for generation tasks with intrinsically diverse targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. pages 445–450.

Jesús Giménez and Lluís Márquez. 2008. Heterogeneous automatic MT evaluation through non-parametric metric combinations. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*. pages 319–326.

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

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 110–119.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches out: Proceedings of the ACL-04 Workshop*. pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. pages 2122–2132.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics* 39(2):267–300.

Ryan Lowe, Michael Noseworthy, Iulian V. Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proc. ACL (to appear). Also presented at ICLR Workshop*.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016a. Natural language inference by tree-based convolution and heuristic matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pages 130–136.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016b. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pages 3349–3358.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. pages 311–318.

Maja Popović, David Vilar, Eleftherios Avramidis, and Aljoscha Burchardt. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. pages 99–103.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. pages 583–593.

Jost Schatzmann, Kallirroi Georgila, and Steve Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proceedings of the 6th Sigdial Workshop on Discourse an Dialogue*. pages 45–54.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069* .

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*. 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*. pages 1577–1586.

Yiping Song, Rui Yan, Xiang Li, Dongyan Zhao, and Ming Zhang. 2016. Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149* .

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 196–205.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation* 24(1):39–50.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Conference of the European Association for Machine Translation*. pages 28–37.

Miloš Stanojević, Amirand Koehn Philipp Kamran, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. pages 256–273.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. pages 3104–3112.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics*. pages 271–280.

Marilyn A Walker, Rebecca Passonneau, and Julie E Boland. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. pages 515–522.

Shuohang Wang and Jing Jiang. 2016. Learning natural language inference with LSTM. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. pages 1442–1451.

Rui Yan, Yiping Song, and Hua Wu. 2016. 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.