The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

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

Chongyang Tao,[1] **Lili Mou,**[2] **Dongyan Zhao,**[1,3] **Rui Yan**[1,3*]
[1]Institute of Computer Science and Technology, Peking University, China
[2]David R. Cheriton School of Computer Science, University of Waterloo
[3]Beijing Institute of Big Data Research, 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, and that RUBER has fair transferability over different datasets.

Introduction

Open-domain human-computer conversation is attracting increasing attention as an established scientific problem (Bickmore and Picard 2005; Bessho, Harada, and Kuniyoshi 2012; Shang, Lu, and Li 2015; Yan et al. 2016; Yao et al. 2017); it also has wide industrial applications like XiaoIce from Microsoft and DuMi from Baidu. Even in a task-oriented dialog (e.g., hotel booking), an open-domain conversational system could be useful in handling unforeseen user utterances.

In existing studies of open-domain conversational systems, however, researchers typically resort to manual annotation to evaluate their models, which is expensive and time-consuming. Hence, 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; Walker, Passonneau, and Boland 2001; Schatzmann, Georgila, and Young 2005). Unfortunately, such evaluation hardly applies to the open domain due to the diversity and uncertainty of utterances: "accuracy" 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, Cherry, and Dolan 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 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 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 of 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.

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

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.

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 correlations with human judgments, and has fair transferability over different open-domain datasets.

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.

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 itself may be universally relevant to all queries (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.[1] The observation implies that a groundtruth alone is insufficient for the evaluation of open-domain dialog systems.

Observation 3. Fortunately, a query itself provides useful information in judging the quality of a reply.[2] Figure 1 plots the average human satisfactory score of a groundtruth reply versus the relevance measure (introduced in the "Referenced Metric" part) 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.

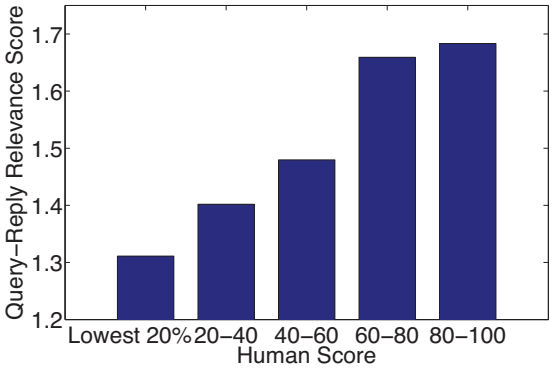


Figure 1: Average query-reply relevance scores versus quantiles of human scores. In other words, we divide human scores (averaged over all annotators) into 5 equal-sized groups, and show the average query-reply relevance score (introduced in the "Referenced Metric" part) of each group.

Methodology

In this section, we design referenced and unreferenced metrics based on the above observations. We will further discuss how they are combined. The overall design methodology of our RUBER metric is shown in Figure 2.

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). As said, they may not be appropriate for open-domain dialog systems (Liu et al. 2016).

We adopt a vector pooling approach that summarizes sentence information by choosing the maximum and minimum values in each dimension of pretrained word embeddings; the closeness of r and $\hat{r}$ is measured by the cosine score. We use such heuristic matching because we assume no groundtruth scores, making it infeasible to train a model with parameters.

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

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

[1] 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.

[2] 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.

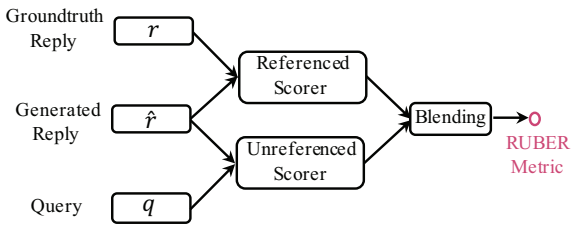


Figure 2: Overview of the RUBER metric.

where $[\cdot]$ indexes a dimension of a vector. Likewise, min pooling yields a vector $\boldsymbol{v}_{\min}$. Since an embedding feature is symmetric in terms of its sign, we cannot tell whether the maximum (positive) value is more important than the minimum (negative) value. We thus use both by concatenating 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

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

Similar heuristics are used in previous work. For example, 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.

Unreferenced Metric

We measure the relatedness between the generated reply $\hat{r}$ and its query q. This metric, denoted as $s_U(q, \hat{r})$, is unreferenced 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 the query q and the reply r is mapped to an embedding; a bidirectional recurrent neural network with gated recurrent units (Bi-GRU RNN) captures information along the word sequence. The forward RNN takes the form

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

where $\boldsymbol{x}_t$ is the embedding of the current input word, and $\boldsymbol{h}^{\rightarrow}_t$ is the hidden state. Likewise, the backward RNN gives hidden states $\boldsymbol{h}^{\leftarrow}_t$. 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

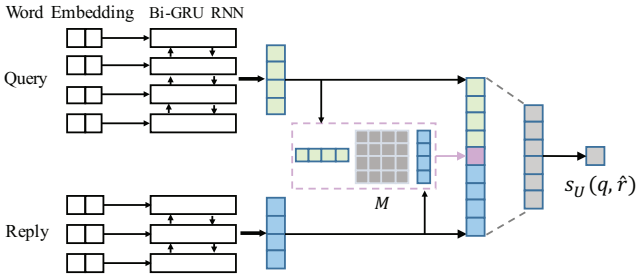


Figure 3: The neural network predicting the unreferenced score.

$\boldsymbol{q}^\top \mathbf{M} \boldsymbol{r}$, where $\mathbf{M}$ is a parameter matrix. Finally, a multilayer 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, Song, and Wu 2016). We may also apply other variants for matching; 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 (set to 0.05 by validation). The training objective is to minimize

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

We train model parameters with *Adam* (Kingma and Ba 2015) with backpropagation.

In previous work, researchers adopt negative sampling for utterance matching (Yan, Song, and Wu 2016; Yan, Zhao, and others 2017). 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 require manual labels for training their metrics (Lowe et al. 2017).

Hybrid Evaluation

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

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

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

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 the experiments, different strategies yield similar results, consistently outperforming baselines.

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. We also compare human-human agreement: "Human (Avg)" refers to average correlation between every two humans, whereas "Human (Max)" refers to the two annotators who are most correlated. Notice that 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.

Experiments

In this section, we evaluate the correlation between our RUBER metric and human annotation, which is the ultimate goal of automatic metrics. We conducted experiments on a Chinese corpus because our the cultural background (as human aspects are deeply involved in this study). However, we shall show the performance of RUBER when it is transferred to different datasets, and we believe our evaluation routine could be potentially applied to different languages.

Setup

We crawled massive data from an online Chinese forum Douban.[3] The training set contains 1,449,218 samples, each of which consists of a query-reply pair. We performed Chinese word segmentation, and obtained Chinese terms as primitive tokens. In the referenced metric, we trained 50-dimensional word2vec embeddings on the Douban dataset. For the unreferenced metric, the dimension of RNN layers was set to 500.

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, Vinyals, and Le 2014) that encodes a query as a vector with an RNN and decodes the vector to a reply with another RNN; the attention mechanism (Bahdanau, Cho, and Bengio 2015) is also applied to enhance query-reply interaction.

We randomly selected 300 samples and invited 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."

Results

Table 2 shows the Pearson and Spearman correlations between the proposed RUBER metric and human scores; also included are various baselines. Pearson and Spearman correlations estimate linear and monotonic correlation, respectively, and are widely used in other research of automatic metrics such as machine translation (Stanojević, Kamran, and Bojar 2015).

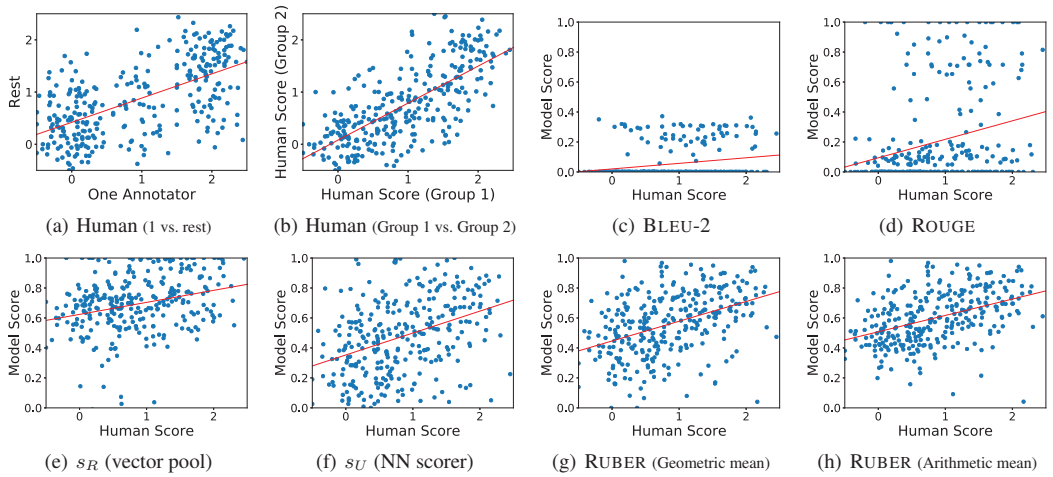
We find that the referenced metric s_R based on embeddings is more correlated with human annotation than existing metrics including both BLEU[4] and ROUGE, which are based on word overlapping information. This implies the groundtruth alone is useful for evaluating a candidate reply. However, exact word 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). Notice that the neural network scorer largely outperforms vector pooling in the unreferenced setting. This is because the cosine measure used in vector pooling 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 ge-

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

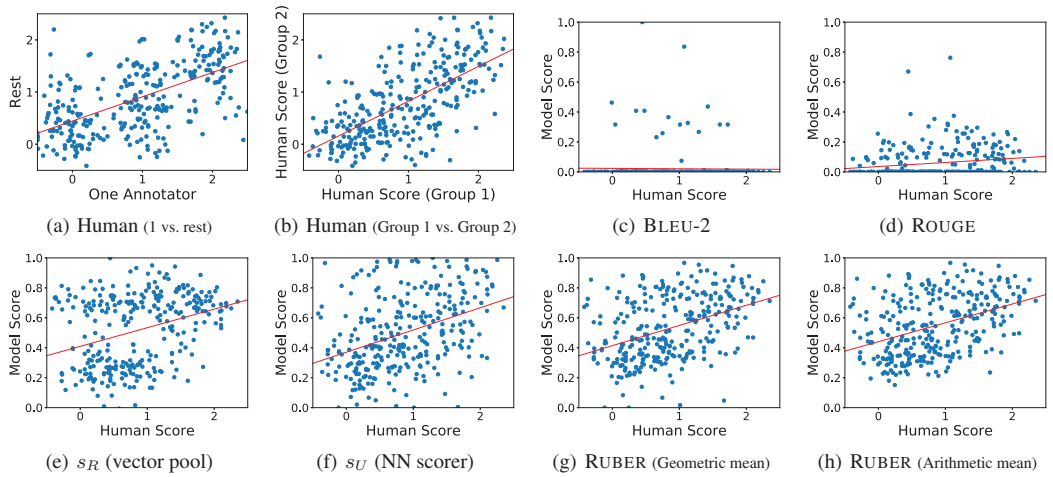
[4] BLEU-n considers n-gram only (instead of a geometric mean of unigram up to n-gram.)



(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).

ometric 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 require human labels; it is more flexible than the existing supervised metric (Lowe et al. 2017), and can be easily adapted to different datasets.

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.

As seen, BLEU and ROUGE scores are zero for most replies, because exact word overlapping occurs very occasionally in short-text conversation; 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.

We would like to clarify more regarding human-human plots. Liu et al. (2016) divide 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. Hence it is not a right statistic to compare with.[5] In our

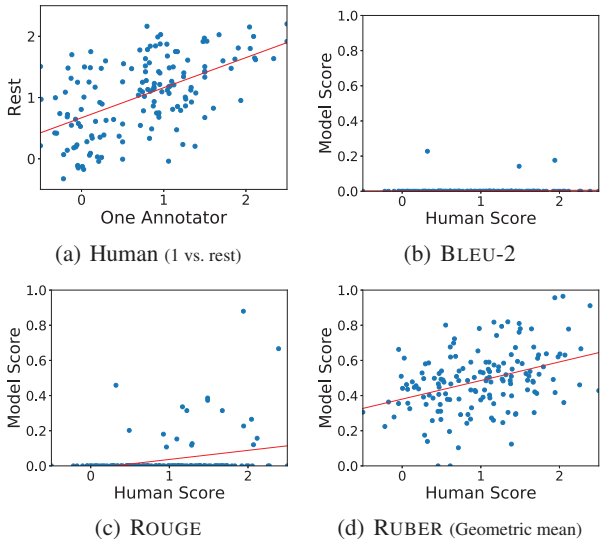
[5]We can image that, in the limit of the annotator number to infinity, Subplots 4b and 5b would become diagonals due to the Law of Large Numbers.

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., $[0, 2]$.

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.

experimental design, 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 medium-correlated human's performance. These qualitative results show our RUBER metric achieves similar correlation to humans.

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 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.

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[6] forum, the topics of which may vary from the previously used Douban corpus. Here we only evaluated the results of the Seq2Seq model (with attention) because of the limit of space and time.

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 previously. Table 4 shows the Pearson and Spearman correlations, 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.

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 (Table 2), showing additional evidence of the effectiveness of our hybrid approach.

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

Related Work

Automatic Evaluation Metrics

Automatic evaluation is crucial to the research of language generation tasks such as dialog systems (Liu et al. 2016), machine translation (Papineni et al. 2002), and text summarization (Lin 2004). The Workshop on Machine Translation (WMT) organizes shared tasks for evaluation metrics (Stanojević, Kamran, and Bojar 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 contents.

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 models. Specia, Raj, and Turchi (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[7] to enhance the correlation with human judgments through supervised learning. However, handcrafted features often require expensive human labor, but do not generalize well. Moreover, these learnable metrics require massive human-annotated scores to learn the model parameters. By contrast, 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 the supervision of human-annotated scores.

Evaluation for Dialog Systems

Dialog systems can also be thought of as a language generation task; several studies adopt BLEU scores to measure the quality of a reply (Li et al. 2015; Song et al. 2016; Tian et al. 2017). However, its effectiveness has been questioned (Callison-Burch, Osborne, and Koehn 2006; Galley et al. 2015). Liu et al. (2016) conduct extensive empirical experiments and show weak correlation of existing metrics (e.g., BLEU, ROUGE, and METEOR) with human judgements for dialog systems. To alleviate the rareness of word overlapping, 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. (2017) and Mou et al. (2016) use entropy to measure the information of generated replies; such metric is independent of the query and groundtruth, and can be easily cheated if used alone. Lowe et al. (2017) propose a neural network-based metric learned in a supervised fashion. By contrast, our approach does not require human-annotated scores.

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

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 considers both the groundtruth and its query. We evaluated RUBER on both retrieval and generative dialog systems. Experiments show that, although unsupervised, RUBER has strong correlation with human annotation, and has fair transferability over different datasets.

Our paper currently focuses on single-turn conversation as a starting point. 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.

Acknowledgments

We thank Yiping Song and Lili Yao for helpful discussions, Dong An, Yaoyuan Zhang, Xinyi Lin et al. for human evaluation, and anonymous reviewers for constructive comments. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001) and the National Science Foundation of China (No. 71672058). Rui Yan was sponsored by the CCF-Tencent Open Research Fund.

References

Albrecht, J., and Hwa, R. 2007. Regression for sentence-level MT evaluation with pseudo references. In *ACL*, 296–303.

Avramidis, E.; Popović, M.; Vilar, D.; and Burchardt, A. 2011. Evaluate with confidence estimation: Machine ranking of translation outputs using grammatical features. In *Proc. Workshop on Statistical Machine Translation*, 65–70.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Banerjee, S., and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

Bessho, F.; Harada, T.; and Kuniyoshi, Y. 2012. Dialog system using real-time crowdsourcing and twitter large-scale corpus. In *SIGDIAL*, 227–231.

Bickmore, T. W., and Picard, R. W. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Computer-Human Interaction* 12(2):293–327.

Bojar, O.; Graham, Y.; Kamran, A.; and Stanojević, M. 2016. Results of the WMT16 metrics shared task. In *Proc. Conf. Machine Translation*, volume 2, 199–231.

Callison-Burch, C.; Osborne, M.; and Koehn, P. 2006. Re-evaluation the role of BLEU in machine translation research. In *EACL*, 249–256.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. Int. Conf. Human Language Tech. Res.*, 138–145.

Forgues, G.; Pineau, J.; Larchevêque, J.-M.; and Tremblay, R. 2014. Bootstrapping dialog systems with word embeddings. In *NIPS Workshop*.

Galley, M.; Brockett, C.; Sordoni, A.; Ji, Y.; Auli, M.; Quirk, C.; Mitchell, M.; Gao, J.; and Dolan, B. 2015. deltaBLEU: a discriminative metric for generation tasks with intrinsically diverse targets. In *ACL-IJCNLP*, 445–450.

Giménez, J., and Márquez, L. 2008. Heterogeneous automatic MT evaluation through non-parametric metric combinations. In *IJCNLP*, 319–326.

Kingma, D., and Ba, J. 2015. Adam: A method for stochastic optimization. *ICLR*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, 110–119.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches out: Proc. ACL-04 Workshop*, 74–81.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*.

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

Lowe, R.; Noseworthy, M.; V. Serban, I.; Angelard-Gontier, N.; Bengio, Y.; and Pineau, J. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *ACL*.

Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *COLING*, 3349–3358.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 311–318.

Popović, M.; Vilar, D.; Avramidis, E.; and Burchardt, A. 2011. Evaluation without references: IBM1 scores as evaluation metrics. In *Proc. Workshop on Statistical Machine Translation*, 99–103.

Ritter, A.; Cherry, C.; and Dolan, W. B. 2011. Data-driven response generation in social media. In *EMNLP*, 583–593.

Schatzmann, J.; Georgila, K.; and Young, S. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *SIGDIAL*, 45–54.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A.; and Bengio, Y. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 3295–3301.

Severyn, A., and Moschitti, A. 2015. Learning to rank short text pairs with convolutional deep neural networks. In *SIGIR*.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *ACL-IJCNLP*.

Song, Y.; Yan, R.; Li, X.; Zhao, D.; and Zhang, M. 2016. Two are better than one: An ensemble of retrieval- and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

Specia, L.; Turchi, M.; Cancedda, N.; Dymetman, M.; and Cristianini, N. 2009. Estimating the sentence-level quality of machine translation systems. In *Proc. Conf. European Assoc. Machine Translation*, 28–37.

Specia, L.; Raj, D.; and Turchi, M. 2010. Machine translation evaluation versus quality estimation. *Machine Translation* 24(1):39–50.

Stanojević, M.; Kamran, Amirand Koehn, P.; and Bojar, O. 2015. Results of the WMT15 metrics shared task. In *Proc. Workshop on Statistical Machine Translation*, 256–273.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Tian, Z.; Yan, R.; Mou, L.; Song, Y.; Feng, Y.; and Zhao, D. 2017. How to make context more useful? An empirical study on context-aware neural conversational models. In *ACL*, volume 2, 231–236.

Walker, M. A.; Litman, D. J.; Kamm, C. A.; and Abella, A. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *EACL*, 271–280.

Walker, M. A.; Passonneau, R.; and Boland, J. E. 2001. Quantitative and qualitative evaluation of darpa communicator spoken dialogue systems. In *ACL*, 515–522.

Yan, R.; Song, Y.; Zhou, X.; and Wu, H. 2016. Shall i be your chat companion?: Towards an online human-computer conversation system. In *CIKM*, 649–658.

Yan, R.; Song, Y.; and Wu, H. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR*, 55–64.

Yan, R.; Zhao, D.; et al. 2017. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *SIGIR*, 685–694.

Yao, L.; Zhang, Y.; Feng, Y.; Zhao, D.; and Yan, R. 2017. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP*, 2180–2189.