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 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 high correlation with human annotation.

1 Introduction

Automatic evaluation is crucial to the research of open-domain human-computer conversation systems. Nowadays, open-domain conversation is attracting increasing attention because of its wide applications (Bickmore and Picard, 2005; Bessho et al., 2012; Shang et al., 2015; Yan et al., 2016). In these 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 automatic 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 conversation systems, researchers occassionally adopt these metrics for evaluation (Ritter et al., 2011; Li et al., 2015; Song et al., 2016). 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 conversation 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 using word-overlapping information (e.g., in BLEU and ROUGE), we measure the similarity by pooling of word embeddings (Forgues et al., 2014); it is more suited to dialog systems due to the diversity of replies.
- A neural network-based scorer measures the relatedness between the generated reply and

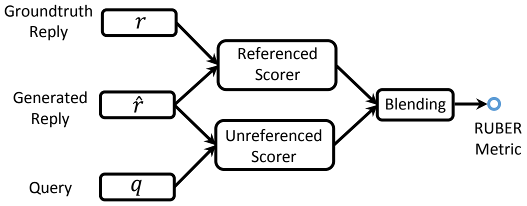

Figure 1: Overview of the RUBER metric.

its query. We observe that 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 labels, 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 plausible topics of the reply; the relatedness between a generated reply and its query then provides additional information. Combining these two aspects with strategies like averaging further improves the performance.

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 against human judgments.

2 Methodology

Figure 1 shows the overall design methodology of our RUBER metric. We introduce the referenced and unreferenced metrics in Subsections 2.1 and 2.2, respectively. Subsection 2.3 discusses how they are combined.

2.1 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). 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

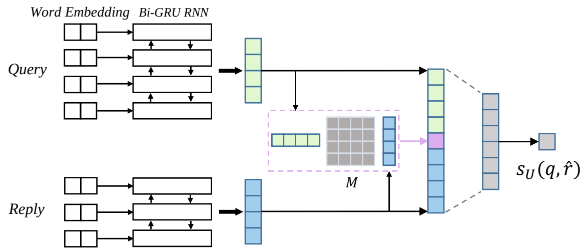

Figure 2: The neural network predicting the unreferenced score.

by the cosine score. We use such heuristic matching because we assume no groundtruth scores, making it unfeasible 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

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

Notice that the pooling approach tends to address uncommon words more than common ones. During unsupervised word embedding learning, common words' vectors are pulled to the origin (near zero) because these words appear frequently in different contexts. Choosing the maximum and minimum values in each dimension extracts more information about uncommon words. Also, such treatment is more robust than vector extrema (Forgues et al., 2014); it chooses either the largest positive or smallest negative value, where features are more vulnerable to signs.

2.2 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 two utterances' similarity, the q-$\hat{r}$ metric in this part involves more semantics. Hence, we empirically design a neural network (Figure 2) 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 sequence. The forward RNN takes the form

$$\boldsymbol{r}_t = \sigma(W_r \boldsymbol{x}_t + U_r \boldsymbol{h}_{t-1}^{\rightarrow} + \boldsymbol{b}_r) \tag{3}$$
$$\boldsymbol{z}_t = \sigma(W_z \boldsymbol{x}_t + U_z \boldsymbol{h}_{t-1}^{\rightarrow} + \boldsymbol{b}_z) \tag{4}$$
$$\tilde{\boldsymbol{h}}_t = \tanh\left(W_h \boldsymbol{x}_t + U_h(\boldsymbol{r}_t \circ \boldsymbol{h}_{t-1}^{\rightarrow}) + \boldsymbol{b}_h\right) \tag{5}$$
$$\boldsymbol{h}_t^{\rightarrow} = (1 - \boldsymbol{z}_t) \circ \boldsymbol{h}_{t-1}^{\rightarrow} + \boldsymbol{z}_t \circ \tilde{\boldsymbol{h}}_t \tag{6}$$

where $\boldsymbol{x}_t$ is the current input, 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., 2016); 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

$$J = \max\left\{0, \Delta - s_U(q, r) + s_U(q, r^-)\right\} \tag{7}$$

All parameters are trained by Adam 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).

2.3 Hybrid Evaluation

We combine the above two metrics by simple heuristics, resulting in a hybrid method RUBER for evaluating 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

$$\tilde{s} = \frac{s - \min(s')}{\max(s') - \min(s')} \tag{8}$$

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 3.2, different strategies yield similar results, consistently outperforming baselines.

3 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 culture background, as human aspects are strongly involved in this paper. We nevertheless believe our evaluation routine could be applied to different languages.

3.1 Setup

We crawled massive data from an online Chinese forum Douban.[1] The training set contains 1,449,218 samples. We performed Chinese word segmentation, and obtained Chinese terms as primitive tokens.

The RUBER metric (along with baselines) is evaluated on two dialog systems. One is a feature-based retrieval-and-reranking system in Song et al. (2016), 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. The other is a sequence-to-sequence neural network (Bahdanau et al., 2015) that encodes a query as a vector with an RNN and decodes the vector

[1] 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)
	VectorPool (s_R)	0.2844(<0.01)	0.3205(<0.01)	0.3434(<0.01)	0.3219(<0.01)
Unreferenced	VectorPool	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 1: Correlation between automatic metrics and human annotation. The p-value is a rough estimation of the probability that an uncorrelated metric produces the result; it does not indicate the degree of correlation.

to a reply with another RNN; the attention mechanism 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.

3.2 Results

Table 1 shows the Pearson and Spearman correlation between the proposed RUBER metric and human scores; also included are various baselines. Pearson and Spearman correlation measures are widely used in other research of automatic metrics such as machine translation (Stanojević et al., 2015).

We find that the reference metric s_R based on mebeddings 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 referenced metric s_U also achieves high correlation, showing that the query alone is also informative[2] and that negative sampling is useful for training evaluation metrics, although it does not requirer human annotation as labels. Our neural network scorer also 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 ($\min$) are 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 Conclusion

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 groundtruth and its query. Experiments show that, although unsupervised, RUBER has strong correlation with human annotation.

[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.

References

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 Representation Learning*.

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. http://aclweb.org/anthology/W05-0909.

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. http://aclweb.org/anthology/W12-1631.

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.

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

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. https://doi.org/10.18653/v1/N16-1014.

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. http://aclweb.org/anthology/W04-1013.

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. https://doi.org/10.18653/v1/D16-1230.

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. https://openreview.net/pdf?id=HJ5PIaseg.

Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2016. 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. https://doi.org/10.18653/v1/P16-2022.

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. http://www.aclweb.org/anthology/P02-1040.

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. http://aclweb.org/anthology/D11-1054.

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.

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*. Beijing, China, pages 1577–1586. https://doi.org/10.3115/v1/P15-1152.

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* .

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the wmt15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. pages 256–273. http://aclweb.org/anthology/W15-3031.

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. http://www.aclweb.org/anthology/P97-1035.

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. http://aclweb.org/anthology/P01-1066.

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. https://doi.org/10.18653/v1/N16-1170.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to rrespond 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.