Are We Evaluating Paraphrase Generation Accurately?

Anonymous ACL submission

Abstract

Paraphrase is a restatement of a text that conveys the same meaning using different expressions. The evaluation of paraphrase generation (PG) is a complex task and currently lacks a complete picture of the criteria and metrics. In this paper, we survey the automatic evaluation metrics and human evaluation criteria of PG evaluation. Base on the survey result, we propose a reference-free automatic toolkit and list clear human evaluation criteria. Moreover, we notice the paraphrases selection in downstream tasks and propose a simple but effective evaluation Filter model. It can fusion multi automatic metrics to fit the human evaluation without any references.

1 Introduction

Paraphrase generation (PG) is a substantial task in the natural language processing (NLP) field. A paraphrase is a restatement of the meaning of a text or passage using other words. An effective PG model is beneficial to many downstream tasks, such as question answering (Yin et al., 2015; Dong et al., 2017; Zhou et al., 2020), duplicate question detection (Shah et al., 2018) and adversarial learning for neural networks (Iyyer et al., 2018).

However, the evaluation of PG is a complex task and currently lacks a complete picture of its criteria and metrics. Different evaluation criteria or automatic metrics often appear in different research. And this makes it is difficult to compare and draw conclusions across papers. To survey the evaluation method, we collect 35 PG research in the past five years. We systematically summary the automatic evaluation metrics and human evaluation criteria, and then make statistics on their frequency.

In **Automatic Evaluation**, we find that there are about 20 metrics have been used in the last five years. The most commonly used metrics, such as BLEU(Papineni et al., 2002) and ROUGE (Lin, 2004), are derived from the evaluation of machine translation (MT) tasks and calculated by referring to a single reference. However, we notice this approach contradicts the expression diversity in paraphrases. As shown in Figure1, *S1* is semantically different from the source but achieves higher BLEU and ROUGE scores. Other sentences which have the same meaning and better expression achieve lower scores. Furthermore, we find the evaluation relying on references violates the definition of paraphrase, both in terms of semantic consistency and diverse expressions. Therefore, we propose to evaluate the predictions of the PG model with reference-free metrics and focus on three aspects: *Relevance* (semantic consistency with source), *Difference* (expression difference with source), *Diversity* (various expressions inside a group of predictions).

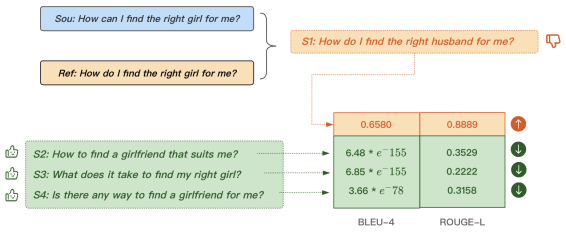

Figure 1: An example for automatic evaluation through BLUE and ROUGE. *S1* with different semantic achieves the highest score. Other lower scores sentences have the same meaning and perfect expression.

In **Human Evaluation**, we summary that the common criteria can be divided into four major categories. However, we find that different work often chooses different categories and the descriptions show high diversity. To unify it, we organize a list set of clear rules based on three aspects: *Relevance* (semantic consistency), *Difference* (expression difference), and *Fluency* (expression fluency).

Our goal is to accurately evaluate the paraphrases, and so that they can serve the downstream tasks. However, for the paraphrases selection in **Downstream Tasks**, it is inadequate to use a sin-

gle automatic metric and unrealistic to use human evaluation. Therefore, we propose a simple but effective Filter model. The Filter can fusion multi-dimensional automatic metrics and get a score similar to the human evaluation without any references.

We summarize our contributions as follows:

- We survey the evaluation of PG in two aspects: automatic evaluation and human evaluation.
- Base on the survey result: (1) In automatic evaluation, we propose multi-dimensional criteria and a reference-free automatic toolkit; (2) In human evaluation, we propose unified criteria and a list set of clear rules.
- We innovatively notice the paraphrase selection in downstream tasks and propose a simple but effective Filter model. It can fusion multi automatic metrics to fit the human evaluation without any references.

2 Automatic Evaluation

Automatic evaluation of PG involves multiple criteria and requires different automatic metrics. However, different metrics make it difficult to compare and draw conclusions across papers.

2.1 Metrics

From the 35 papers in the past five years, we statistics the metrics that have been used for the evaluation on the Quora Duplicated Questions[1] and summary them as follows:

Expression Consistence with references is the most common PG automatic criterion. The widely-used metrics include: (1) BLEU(Papineni et al., 2002) is a common metric that uses N-gram matching rules; (2) ROUGE-n and ROUGE-L(Lin, 2004) are the recall-based evaluation metrics; (3) METEOR(Denkowski and Lavie, 2014) can measure partial semantic equivalents; (4) TER(Snover et al., 2006) measures the amount of editing.

Semantic Relevance with source is valued by some PG tasks. The widely-used metrics include: (1) Bertscore(Zhang et al., 2019) computes a similarity score for each token; (2) Setence-BERT (SBERT)(Reimers and Gurevych, 2019) computes the cosine similarity of sentence-level embeddings; (3) Embedding Average Cosine Similarity (EACS) and Greedy Matching Score (GMS)(Sharma et al., 2017) measure the similarity based on the cosine similarity of embeddings on word and sentence levels; (4) Paraphrase Detection score (PDS)(Kumar et al., 2020) is a classifiers model trained on the task of Paraphrase Detection.

Expression Difference with the source appears frequently in recent PG work. The widely-used metric is: (1) self-BLEU(Cao and Wan, 2020) calculate BLEU between the prediction and source.

Diversity inside the set of predictions is used by the multi-output PG tasks. The widely-used metrics include: (1) Dist-n(Li et al., 2015) measures the number of distinct n-grams within the predictions; (2) mBLEU(Fan et al., 2018) computes the dissimilarity of BLEU scores within the predictions; (3) Pairwise-BLEU(Cao and Wan, 2020) evaluates the average difference between the k predictions from the same source; (4) self-BLEU(Zhu et al., 2018) evaluate the BLEU within the predictions.

Syntax Structure consistence is evaluated by some PG tasks. (1) TED-E and TED-R(Kumar et al., 2020) evaluate the syntactic transfer using Tree-edit distance(Zhang and Shasha, 1989) between the parse trees of the predictions with the syntactic exemplars or with the references; (2) Parse Tree Similarity(Iyyer et al., 2018) calculates the top two levels of parse tree similarity among the predictions; (3) Syntactic Tree (ST) Edit Distance(Chen et al., 2019) computes the Tree-edit distance between parse trees after removing word tokens.

Other. Besides the above criteria, there are some metrics for fluency and cross-evaluation: (1) Negative Likelihood (NLL)(Miao et al., 2019) is used to measure the fluency of the predictions; (2) iBLEU(Sun and Zhou, 2012) penalizes the similarity of the predictions with the source besides the expression relevance to the references; (3) BERT-iBLEU(Niu et al., 2020) encourages semantic closeness while penalizing surface-form.

2.2 Statistics

As summarized above, there are about 20 automatic metrics have been used in the recent PG work. We make statistics on the times they have been used. Table 1 shows the most commonly metrics.

	BLEU	**ROUGE-n**	**MET**	**ROUGE-L**
Times	34	23	18	11
	iBLEU	**self-BLEU**	**TER**	**BERTscore**
Times	9	7	6	3

Table 1: The Top-8 commonly used automatic metrics and their occurrences times. MET is METEOR.

[1] https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

The metrics such as BLEU, ROUGE, METEOR, which calculate by reference and derive from MT task, are the most common. We explore whether these metrics are equally applicable in PG tasks.

2.3 Experiment

We focus on the metrics relying on references and conduct the experiment on the Quora Duplicated Questions dataset. Compare with the Baselines: *Source*: taking the source sentence as results; *Synonym*: replacing a random word in the source sentence with its synonym; *State-of-the-art PG Approaches*: DNPG (Li et al., 2019), LBOW-Topk (Fu et al., 2020) and IANet+S (Su et al., 2021).

	BLEU-4 (%) ↑	**Rouge-L (%) ↑**
Source	34.41	63.14
Synonym	25.03	56.29
DNPG	25.03	-
LBOW-Topk	26.17	56.43
IANet+S	27.09	58.01

Table 2: The automatic evaluation results of the metrics that referring to the reference on the Quora dataset.

Table 2 show the results. Compare with the recently excellent PG approaches, completely copying the source achieves a better score and *Synonym* achieves comparable scores. However, as we know, these are worthless to the downstream tasks. So it is incomplete to use the metrics which refer to the reference for the PG evaluation.

2.4 Reference-Free Toolkit

To avoid the drawbacks of relying on references evaluation and fit the PG requirements, we propose a reference-free evaluation toolkit that considers three aspects: *Relevance*, *Difference*, and *Diversity*. The evaluation toolkit includes a series of automatic evaluation metrics:

Semantic Relevance: BERTscore

Expression Difference: self-BLEU

Group Diversity: group-BLEU (self-BLEU inside the set of predictions)

3 Human Evaluation

Human evaluation plays an important role in PG evaluation. In the 35 PG work, 26 do the human evaluation. However, different PG research often chooses different criteria categories and shows high diversity in the descriptions for each category.

3.1 Criteria

From 22 papers that have described the detail of human evaluation criteria, we summary the criteria and their keywords.

Fluency is an important criteria for human evaluation. The described keywords of this criterion are fluency, readability, and plausibility.

Relevance in semantic is also a consideration in human evaluation like automatic evaluation. The described keywords of this criterion are relevance, semantic, accuracy, consistency, coherence, equivalence, and fidelity.

Difference in expression is included in some human evaluation criteria. The described keywords of this criterion are difference, diversity, surface dissimilarity, and variability.

Diversity is evaluated by a few PG work with multi-output. The described keywords of this criterion are diversity and variability.

3.2 Statistics

We make statistics on the times of these criteria appear, Table 3 shows the result.

	Relevance	**Fluency**	**Difference**	**Diversity**
Times	22	19	11	1

Table 3: In the 22 research, the occurrences times of the human evaluation criteria.

Relevance, Fluency, and Difference are the most common criteria in the human evaluation and it is reasonable for the downstream tasks as well. However, the different selection and descriptions of the human evaluation criteria bring inconvenience to compare across research. We propose to use unified criteria with the same keywords.

3.3 Unified Criteria

As claimed in (Howcroft et al., 2020), human evaluation in natural language generation presents confusion and is in urgent need of standard methods. The human evaluation in PG tasks needs unified criteria. We propose to perform human evaluation and grade division from three aspects: ***Relevance*** (Semantic Consistency), ***Fluency*** (Expression Fluency), and ***Difference*** (Expression Difference). Since human evaluation objects are separate paraphrase pairs, we use fluency instead of diversity, which is the important criterion in automatic evaluation. The specific criteria is shown in Table 4.

Score	Criteria			Example
	Relevance	*Fluency*	*Difference*	
0	-	-	×	How can I find the girl for me?
1	×	×	-	How do you feed your cat?
2	O	×	-	How can you find your cat?
3	√	√	O	How can I find the girlfriend for me?
4	√	√	O	How do I find my right girl?
5	√	√	√	Is there any way to find my right girl?

Table 4: Human Evaluation Criteria. The example is for the source: "How can I find the girl for me?". "O" means partly meets the criteria. The difference between 3 and 4 points is the range of differences: words or structures.

4 Evaluation in Downstream Tasks

After automatic evaluation and human evaluation, our ultimate goal is the paraphrases can serve downstream tasks. For the selection of paraphrases in downstream tasks, it is inadequate to use a single automated metric and unrealistic to use human evaluation. Therefore, we propose a reference-free Filter model and it can fusion multi-dimensional automatic metrics and fit the human evaluation.

4.1 Model

For each paraphrase pairs $\langle X, \widehat{Y} \rangle$, we pick out a human evaluation score H according to the criteria in Section 3.3 and a set of reference-free features $A = [a_1, a_2, ..., a_n]$. For A, based on the criteria in Section 2.4, we expand part automatic metrics to capture more features and delete the diversity metrics for the separately paraphrase pair. Specifically, we pick out BERTscore, self-BLEU, self-ROUGE-n, self-ROUGE-L, Edit Distance, and Jaccard Distance. Our objective is to build a mapping function $f : A \rightarrow H$. We apply an XGBoost(Chen and Guestrin, 2016) model and use additive functions to predict the output.

4.2 Experiment

We construct a dataset to train the Filter. It contains 13,335 paraphrases pairs by mapping from the automatic scores to human evaluation scores. To demonstrate the effectiveness of the Filter, we conduct data augmentation experiments on the Text Classification task through paraphrase generation.

4.2.1 Setup

Dataset. We evaluate the Filter on StackOverflow [2] (Xu et al., 2015) dataset. We randomly select 500, 1000, 3000 data for the training set, 1000 data for the development set, and 2000 data for the test set.

[2] https://github.com/jacoxu/StackOverflow

Data Augmentstion. We generate paraphrases by a sequence-to-sequence model for each original sentence with a beam size of 5. We select the unfiltered predictions or the Top-3 predictions which score by the Filter and pair the results with the tags of the original sentence for data augmentation. BERT (Devlin et al., 2019) is used as a multi-classification baseline for testing.

4.2.2 Results

As shown in Table 5, after selecting by our Filter, not only the amount of data is reduced, but it also promotes accuracy. This shows the effectiveness of the Filter in the downstream task.

Data Size	Base (%)	Baseline	**Accuracy (%)**
500	*77.25*	BERT	77.65
		$\text{BERT}_{\text{+Filter}}$	**77.70**
1000	*79.75*	BERT	80.25
		$\text{BERT}_{\text{+Filter}}$	**81.10**
3000	*83.10*	BERT	83.40
		$\text{BERT}_{\text{+Filter}}$	**83.95**

Table 5: The accuracy of Text Classification task. Base represents the result without data augmentation.

5 Conclusion

In this paper, we analyze the evaluation of PG from three aspects: automatic evaluation, human evaluation, and downstream tasks. First, we survey the PG evaluation in the automatic metrics and human evaluation. Base on the survey result, in automatic evaluation, we propose a set of multi-dimensional criteria and create a reference-free toolkit. It can avoid the limitations of the common evaluation method relying on reference. For human evaluation, we propose unified criteria and a list set of clear rules. Moreover, to make up for the gap of the PG evaluation in downstream tasks, we propose a simple but effective Filter model. It can fusion multi automatic metrics and fit human evaluation to enhancement in downstream tasks.

References

Yue Cao and Xiaojun Wan. 2020. Divgan: Towards diverse paraphrase generation via diversified generative adversarial network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2411–2421.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. *arXiv preprint arXiv:1906.00565*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. *arXiv preprint arXiv:1708.06022*.

Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018. A question type driven framework to diversify visual question generation. In *IJCAI*, pages 4048–4054.

Yao Fu, Yansong Feng, and John P Cunningham. 2020. Paraphrase generation with latent bag of words. *arXiv preprint arXiv:2001.01941*.

David M Howcroft, Anja Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345.

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

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. 2019. Decomposable neural paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3403–3414.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2020. Unsupervised paraphrasing with pretrained language models. *arXiv preprint arXiv:2010.12885*.

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 of the Association for Computational Linguistics*, pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Darsh J Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. 2018. Adversarial domain adaptation for duplicate question detection. *arXiv preprint arXiv:1809.02255*.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Matthew G. Snover, B. Dorr, R. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA*.

Yixuan Su, David Vandyke, Simon Baker, Yan Wang, and Nigel Collier. 2021. Keep the primary, rewrite the secondary: A two-stage approach for paraphrase generation. *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, pages 1–6.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual smt system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015.

Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69.

Pengcheng Yin, Nan Duan, Ben Kao, Junwei Bao, and Ming Zhou. 2015. Answering questions with complex semantic constraints on open knowledge bases. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1301–1310.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM journal on computing*, 18(6):1245–1262.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Huiyang Zhou, Haoyan Liu, Zhao Yan, Yunbo Cao, and Zhoujun Li. 2020. Larq: Learning to ask and rewrite questions for community question answering. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 318–330. Springer.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.