Perplexity from PLM Is Unreliable for Evaluating Text Quality

Yequan Wang\textsuperscript{1}, Jiawen Deng\textsuperscript{2}, Aixin Sun\textsuperscript{3}, Xuying Meng\textsuperscript{4}
\textsuperscript{1}Beijing Academy of Artificial Intelligence, Beijing, China
\textsuperscript{2}CoAI Group, DCST, IAI, BNRIST, Tsinghua University, Beijing, China
\textsuperscript{3}School of Computer Science and Engineering, Nanyang Technological University, Singapore
\textsuperscript{4}Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
tshwangyequan@gmail.com, dengjw2021@mail.tsinghua.edu.cn, axsun@ntu.edu.sg, mengxuying@ict.ac.cn

Abstract

Recently, amounts of works utilize perplexity (PPL) to evaluate the quality of the generated text. They suppose that if the value of PPL is smaller, the quality (i.e., fluency) of the text to be evaluated is better. However, we find that the PPL referee is unqualified and it cannot evaluate the generated text fairly for the following reasons: (i) The PPL of short text is larger than long text, which goes against common sense, (ii) The repeated text span could damage the performance of PPL, and (iii) The punctuation marks could affect the performance of PPL heavily. Experiments show that the PPL is unreliable for evaluating the quality of given text. Last, we discuss the key problems with evaluating text quality using language models.

1 Introduction

With the development of natural language processing, generation tasks (i.e., machine translation (Tu et al., 2016; Zhang et al., 2021), question answer (Duan et al., 2017), generation-based dialog system (Tu et al., 2022), et al.) have seen tremendous growth and progress. How to evaluate the quality of generated text fairly is important and urgent.\textsuperscript{1}

There are various statistical metrics including word-based metrics (i.e., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)), character-based metrics (i.e., chrF (Popovic, 2015)), embedding-based metrics (i.e., Vector Extrema (Forgues et al., 2014), Greedy Matching (Rus and Lintean, 2012)) and their variants to evaluate the performance of generation methods. BLEU is capable of reflecting the ratio of overlapping \textit{n}-grams to the total \textit{n}-grams, which denotes the precision-based measure. Different from BLEU, ROUGE and its variants could reflect the match at the recall-based measure (Sai et al., 2023). Vector Extrema can prioritize informative words by taking the extreme value along each dimension. However, all of those metrics are based on statistics or simple vector representation-based statistics, they cannot evaluate the creativity, diversity, and complexity of generated text even when people try to express the same meaning.

Besides, there are other evaluation metrics (Shimanaka et al., 2019) based on pre-trained language models (PLMs), among which PPL is widely used to quantify how well language models learn natural language (Meister and Cotterell, 2021a). It uses the cross-entropy of the probability distribution of predicted token and ground-truth token of text to evaluate the fluency of generated text. Here, the PLM could directly use mainstream models, such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019) et al., so existing studies think that the PPL is able to reflect the fluency of generated text.

However, we find that the current PPL cannot fairly evaluate the text quality (i.e., fluency) when meeting the following scenarios. (i) The texts to be evaluated have different lengths (Meister and Cotterell, 2021b). In fact, text quality is not strictly related to length. However, we find that the PPL is sensitive to text length, e.g., the PPL of short text is larger than long text. (ii) The texts to be evaluated have some repeated span(s). Of course, sometimes creators use repeated text span(s) to express emphasis et al.. However, PPL cannot distinguish between the right emphasis and abnormal repetition, and always foolishly assigns lower scores to text that is not fluent but has repeated spans. (iii) The texts to be evaluated are sensitive to punctuation marks. For example, we have two texts, the former ends with punctuation, and the latter deletes the last punctuation. In theory, the qualified metric should compute the same or similar value. However, there is a significant difference between the PPL values of those two texts.

Our paper summarizes the above vital findings of

\textsuperscript{1}Indicates equal contribution
\textsuperscript{1}We consider task agnostic only in this paper.
PPL failure. Further, we shed light on future potentially better evaluation metrics that should consider the following aspects to ensure fairness: (i) not sensitive to length; (2) sensitive to common mistakes, e.g., unusual repeated text; (3) not sensitive to reasonable punctuation increases or decreases. Specifically, considering text fluency while penalizing duplicate text spans and reducing attention to non-significant punctuation is a good router.

To the best of our knowledge, this is the first attempt to systematically reveal that PPL is not up to the task of text generation evaluation. And we detail the main reason why PPL is unreliable is that PPL cannot handle the typical cases well. Last, we suggest a possible better evaluation direction.

2 Approach

In this section, we briefly introduce the method used to evaluate the performance of PPL. We follow the mainstream methods using GPT-2 (Radford et al., 2019) as the pre-trained language model to calculate PPL.

Given one sentence \([w_1, w_2, \ldots, w_n]\), we generate the token sequence \(s = [t_1, t_2, \ldots, t_m]\) with \(m\) size. Follow the settings of existing studies, we use GPT2-large to compute the PPL:

\[
P_1, P_2, \ldots, P_m = \text{GPT-2}([t_1, t_2, \ldots, t_m]),
\]

where \(P_i\) denotes the predicted probability of \(i\)-th token. Using the cross entropy of each token, we could calculate the PPL of this input sentence through:

\[
PPL(s) = \exp \left\{ \frac{1}{m} \sum_{i=1}^{m} \text{cross-entropy}(t_i, P_i) \right\}.
\]

Here, \(PPL(s)\) is the perplexity value of the input sentence \(s\).

2.1 The PPL of text with different length

Is PPL sensitive to text length? To evaluate the performance of PPL when meeting texts of different lengths, we design this simulation.

Given sentence \(s = [t_1, t_2, \ldots, t_m]\), we use Equation 1 and 2 to calculate the PPL value. In human cognition, in general, the length of text has little relationship with the quality of the generated text. In other words, the PPL value should remain stable and not change with the length of the text.

2.2 The PPL of text with unusual repeated text span

Does PPL make sense when the text lengths are at the same level? We know that if a generative model does not perform well, it may cause duplication in the late generation. To evaluate the performance of PPL, we design the following rule to imitate this kind of situation.

Given sentence \(s = [t_1, t_2, \ldots, t_m]\), we repeat the last \(q\) tokens to generate the sentence \(s_1\):

\[
s_1 = [s, s_l \otimes e_k],
\]

where \(s_l = [t_{m-q+1}, \ldots, t_m]\) denotes the latter text with \(q\) tokens. The operation \(\otimes\) here means: \(s_l \otimes e_k = [s_l, s_l, \ldots, s_l]\), that is, the operation repeatedly concatenates \(s_l\) \(k\) times, where \(e_k\) is a vector with \(k\) 1s (Wang et al., 2016). Then we could use Equation 1 and 2 to compute the PPL of new sentence \(s_1\) with \(m + q \ast k\) size.

2.3 The PPL of text affected by punctuation marks

Is PPL sensitive to punctuation? We know that punctuation marks are important for human reading. So what will happen if we remove all punctuation marks or simply remove the last punctuation mark if it is located at the end of the text? To this end, we design the following rules to imitate the situation.

(I) Remove all punctuation marks. To generate the new text \(s_2\) without punctuation marks, we use the rule through

\[
s_2 = [\ldots, t_i, \ldots],
\]

where \(t_i\) is not a punctuation mark. Otherwise, it will be deleted. Then Equation 1 and 2 are used to compute the PPL of \(s_2\).

(II) Remove the last punctuation mark located at the end of the text. To generate the new text \(s_3\) without the last punctuation mark, we use the rule through

\[
s_3 = \begin{cases} [t_1, t_2, \ldots, t_m], & \text{if } t_m \neq \text{punc.} \\ [t_1, t_2, \ldots, t_{m-1}], & \text{otherwise.} \end{cases}
\]

Here, we do not consider if a text ends with two continuous punctuation marks because the testing dataset from GPT-2 used in our experiments has little data like this. Similarly, Equation 1 and 2 are used to calculate the PPL value.
3 Experiment

We now evaluate the performance of PPL to prove that the PPL is not qualified for generated text evaluation. In the experiments, we use the testing data from the test split of WikiText-2 dataset \(^2\) and take GPT2-large as the PLM. To prevent the effect of extreme lengths, we filter sentences with less than 3 words. There are 2,786 texts left, with maximum, minimum, and average lengths of 481, 3, and 86.52 words, respectively.

As we mentioned in Section 1, PPL performs unreliably for mainly three reasons, including (i) text length, (ii) repeated text span, and (iii) punctuation mark. Next, we experiment on each scene.

3.1 Evaluation on text length

We know that text quality is not very related to text length. If the evaluation metric is sensitive to text length, then the value of this metric is not convincing. So we conduct the experiment on text length to find the inconsequence of PPL. Figure 1 details the changing trend of PPL as the text length changes.

**Observations.** We find that the PPL value gets smaller as the text length gets longer. Obviously, it does not fit our common sense. Text quality (i.e., fluency) has little relation to text length. Even when the length of the text becomes longer, the text is more difficult to understand, resulting in higher PPL. However, the experimental result reveals that the PPL has an opposite trend when meeting the text with different lengths. More importantly, the length of the generated text is variable so the problem is more serious.

2\url{https://huggingface.co/datasets/wikitext}

| q | k | ppl_avg | ppl_std | Len_avg | Increase_Rate |
|---|---|---------|---------|---------|---------------|
| 0 | 0 | 411.99  | 3546.23 | 86.52   | –             |
| 1 | 1 | 235.39  | 1333.33 | 87.52   | 70.17         |
| 1 | 3 | 87.78   | 250.82  | 89.52   | 24.59         |
| 1 | 9 | 36.48   | 363.83  | 95.52   | 0.54          |
| 1 | 12| 34.93   | 583.00  | 98.52   | 0.36          |
| 5 | 1 | 50.21   | 70.96   | 91.48   | 39.16         |
| 5 | 3 | 21.85   | 13.09   | 101.41  | 0.86          |
| 5 | 9 | 10.37   | 6.97    | 131.18  | 0.00          |
| 5 | 12| 8.35    | 5.75    | 146.07  | 0.00          |
| 10| 1 | 35.74   | 59.70   | 95.61   | 2.40          |
| 10| 3 | 14.43   | 8.22    | 113.79  | 0.00          |
| 10| 9 | 6.13    | 4.04    | 168.33  | 0.00          |
| 10| 12| 4.84    | 3.15    | 195.60  | 0.00          |

3.2 Evaluation on the number of text repeated time

A big problem for the generation model is the unusual repetition of generated text. For example, a dialog system may generate “Natural language processing is a subfield of linguistics, computer science science science...” to answer “Can you tell me what natural language processing is?”. The repeated text span damages the semantics of the text. To this end, the qualified referee should have the ability to be aware of this unusual and give them a bad performance or review. To mimic the generative model realistically, we do experiments where the last N tokens of the text are repeated K times. Table 1 details the result. Especially, Figure 2 gives the experiment result of repeating the entire text 2, 3, and 4 times.

**Observations.** Both Figure 2 and Table 1 tell us that the value of PPL gets smaller as the number of text repetition times increases. It is not normal. We know that the unusual text repetition damages the semantic structure of text so the corresponding PPL should be larger. However, the PPL performs opposite direction. This experiment proves that the PPL could not be competent when meeting the unusual repeated text.

Of course, language sometimes uses text repetition to express emotions such as emphasis, anger et al.. When meeting this kind of text, we do not want the metric changes a lot. However, the experimental result tells us that PPL value always drops sharply in case of replication. This is because the
used PLM model (i.e., GPT-2) is trained by lots of language text, and such special text representations are relatively rare.

### 3.3 Evaluation on punctuation marks

Punctuation uses spacing, conventional signs (called punctuation marks) et al. to help understanding and correct reading of the written text. If we read the original version of Shakespeare’s poem or Chinese ancient poetry, we find that it is not easy. One important reason is the lack of punctuation marks. A good metric is not seriously affected by the inessential punctuation or the reasonable absence of punctuation marks. We conduct experiments to study the effect of punctuation. Table 2 shows the statistical results of PPL affected by punctuation marks. Figure 3 depicts the distribution of removing all punctuation marks or the ones located at the end of the text.

**Observations.** Table 2 shows that over 13% sentences have an unreasonable tendency when deleting all punctuation marks. Unfortunately and importantly, both Table 2 and Figure 3 show that if we add or remove a punctuation mark that does not affect the semantics of language, the PPL value has a significant change. This is what we don’t want to see. But it happens in the current PPL.

### 4 Discussion and Future Work

In the above experiments, we first study the effect of text length. The results tell us that the PPL value is sensitive to the text length. Obviously, this problem damages the fairness of PPL. Meanwhile, the good news is the PPL tends to be stable if the text length is greater than 100. Then, we conduct the next experiment to study the effect of unusual repeated text span. However, PPL not only fails to remain stable, but also deviates negatively from the desired direction. Lastly, we study the effect of punctuation marks. Experimental results show that PPL still cannot achieve the desired effect. To give an intuitive description, the case study shown in Section A.1 is conducted.

The above experiments show that the PPL from PLM is unreliable for evaluating text quality. We have the following suggestions to improve the evaluation. We argue that a fair metric should consider the impacts of text length, unusual text repetition, and punctuation perturbation. Besides, diversity is an important indicator of text quality. We think the fluctuations of PPL due to repetition problems could be mitigated with text diversity, e.g., penalizing the scoring based on the frequency of n-gram occurrences, and even penalizing consecutive occurrences of the same expression. Therefore, it is very important that a new metric could consider both PPL and diversity in a way like F1 score. For text quality, the correct use of punctuation is undoubtedly essential. But in the case of lower text quality, a good evaluation metric should focus more on word coherence while reducing attention to non-significant punctuation (e.g., mark "." at the end position) that will not interfere with one’s comprehension. Considering these factors, an reliable evaluation metric is expected to be developed to assess the quality of a text.
5 Conclusion

In this paper, we study the reliability of PPL. First, we find that the PPL is sensitive to text length, which makes it incapable of evaluating text quality. Second, we find that it cannot handle the common problem of unusual text duplication. Last, we find that it could be affected by punctuation marks. To this end, we suggest that develop a new evaluation metric considering both PPL and diversity.

Limitations

In this paper, we present the typical scenarios that PPL could not perform fairly when evaluating the quality of generated text. That is why we need to carefully consider whether to use this metric. Further, we summarize the vital problem of PPL failure, then give suggestions to improve the validity and fairness of the evaluation. However, the limitation is that we do not give a detailed solution. And in the future, we will continue to study it.

Acknowledgments

This work was supported by the National Key R&D Program of China (2020AAA0105200) and the National Science Foundation of China (NSFC No. 62106249, 61902382, 61972381).

References

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, NAACL-HLT 2019. Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pages 866–874. Association for Computational Linguistics.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In Nips, modern machine learning and natural language processing workshop, volume 2, page 168.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Clara Meister and Ryan Cotterell. 2021a. Language model evaluation beyond perplexity. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 5328–5339. Association for Computational Linguistics.

Clara Meister and Ryan Cotterell. 2021b. Language model evaluation beyond perplexity. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5328–5339, Online. Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA, pages 311–318. ACL.

Maja Popović. 2015. chrf: character n-gram f-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal, pages 392–395. The Association for Computer Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Vasile Rus and Mihai C. Linteau. 2012. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, BEA@NAACL-HLT 2012, June 7, 2012, Montréal, Canada, pages 157–162. The Association for Computer Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. A survey of evaluation metrics used for NLG systems. ACM Comput. Surv., 55(2):26:1–26:39.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with BERT regressor. CoRR, abs/1907.12679.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Jirong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long.
Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers. The Association for Computer Linguistics.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016, pages 606–615. The Association for Computational Linguistics.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, and Yang Liu. 2021. Neural machine translation with explicit phrase alignment. IEEE ACM Trans. Audio Speech Lang. Process., 29:1001–1010.
Table 3: Cases of unfair PPL when meeting texts with varying lengths. Combine denotes the original text from the Wikitext-2 dataset. Split1 and Split2 are the sub-sentences obtained by the slicing operation.

| Text                                           | Len. | PPL  |
|------------------------------------------------|------|------|
| Split1: Over the course of his reign, Nero often made rulings that pleased the lower class. | 17   | 77.3 |
| Split2: Nero was criticized as being obsessed with personal popularity. | 10   | 187.9|
| Combine: Over the course ... the lower class. Nero was criticized ... with personal popularity. | 27   | 60.6 |
| Split1: Lesnar appears in the video games WWE SmackDown! Shut Your Mouth, WWE SmackDown! | 16   | 26.3 |
| Split2: Here Comes the Pain, Madden NFL 06, UFC 2009 Undisputed, UFC Undisputed 2010, WWE '12, WWE '13, WWE 2K14, WWE 2K15, WWE 2K16, and WWE 2K17. | 38   | 10.4 |
| Combine: Lesnar appears in ..., WWE SmackDown! Here Comes the Pain, ... and WWE 2K17. | 54   | 8.7  |
| Split1: The Great Fire of Rome erupted on the night of 18 July to 19 July 64. | 17   | 62.8 |
| Split2: The fire started at the southeastern end of the Circus Maximus in shops selling flammable goods. | 17   | 65.7 |
| Combine: The Great Fire..., 19 July 64. The fire started... selling flammable goods. | 34   | 31.9 |
| Split1: This isn’t acceptable to get to where we want to go. But what does that really mean? | 20   | 40.9 |
| Split2: It’s not just get better defensively, it is, if we give up 3 less baskets a game, then we will be at 40 percent field goal percentage defense which will be top 20 in the country | 40   | 44.3 |
| Combine: This isn’t acceptable ..., that really mean? It’s not just ... in the country | 60   | 32.0 |

Table 4: Cases for unreliable PPL. "q, k" denotes repeat the last q tokens k times. "2-Times" denotes repeat the whole text. "w/o Punc." and "w/o Last Punc." denote removing all punctuations and the last punctuation, respectively.

| ID | Text                                                                 | Len. | Orig. | q1, k3 | q1, k9 | q5, k3 | 2-Times w/o Punc. | w/o Last Punc. |
|----|----------------------------------------------------------------------|------|-------|--------|--------|--------|--------------------|----------------|
| 1  | To use absolutely no word that does not contribute to the presentation. | 13   | 140.1 | 64.8   | 24.0   | 20.5   | 17.0   | 106.6           | 106.6          |
| 2  | Coastal service and riverine vessels, including 'floating batteries' and 'monitors' | 14   | 321.4 | 255.1  | 41.2   | 53.6   | 31.1   | 282.3           | 235.0          |
| 3  | Plans, mobilization, and escalating violence = = = = = = = = = = | 11   | 952.6 | 370.4  | 60.2   | 32.3   | 55.8   | 1,690.1        | 1,870.9        |
| 4  | Loans to Peterborough and Molde = = = = = = = = = = = = | 11   | 295.6 | 110.5  | 31.3   | 19.9   | 22.7   | 216.6          | 488.2          |
| 5  | Gregzilla (Greg Harrison) — guitar (2015 – current) | 12   | 711.1 | 519.5  | 117.5  | 39.5   | 41.5   | 50,441.8       | 1,210.7        |
| 6  | Above the cut @-@ off frequency the image impedance is imaginary. | 12   | 1,075.6 | 505.3 | 103.7  | 58.3   | 57.5   | 500.1           | 1,131.9        |

A Appendix

A.1 Case Study

In order to have an intuitive feeling, we selected data from the testing set as a case study. Table 3 and Table 4 report the detail, and reveal that the PPL metric is unreliable.

Table 3 shows that long texts are more likely to get lower values. But when it is sliced into multiple sub-sentences, the PPL of each piece always rise substantially, even if they use the same word sequence as the original long texts.

As reported in Table 4, PPL is deeply perturbed by unusual text span duplication. It shows the trend that PPL decrease is greater as the length of the duplicated text increases (Orig. -> q1, k3 -> q1, k9 -> q5, k3). This indicates that low-quality text can easily fool the metric by a large number of repetitions and achieve lower PPL values, making PPL metric less trustworthy.

PPL is also too sensitive to punctuation. For example, after removing the last punctuation in the third sentence, the PPL rises to 196.4% of the original text (increased from 952.6 to 1870.9). And even the removal of punctuation of little significance (e.g. full stop) does not escape the disaster of PPL fluctuations. For example, in the first sentence, ppl drops to 76.1% of the original sentence (decrease from 140.1 to 106.6) after removing the last mark.

Perturbations from unusual text span and punctuations are common problems in generative models. However, the above-mentioned cases tell us that these problems can easily break the PPL metric, cause large fluctuations and make it difficult to score the text quality fairly.