Rethinking and Refining the *Distinct* Metric

Anonymous ACL submission

Abstract

Distinct is a widely used automatic metric for evaluating the diversity of language generation tasks. However, we observe that the original approach to calculating distinct scores has evident biases that tend to add higher penalties to longer sequences. In this paper, we refine the calculation of distinct scores by re-scaling the number of distinct tokens based on its expectation. We provide both empirical and theoretical evidence to show that our method effectively removes the biases exhibited in the original distinct score. Further analyses also demonstrate that the refined score correlates better with human evaluations.

1 Introduction

The diversity of generated texts is an important evaluation aspect for dialogue generation models since most neural dialogue models tend to produce general and trivial responses (like "I don’t know" or "Me too") (Li et al., 2016; Zhao et al., 2017). Several metrics have been proposed to evaluate the text diversity, and the *Distinct* score proposed by Li et al. (2016) is the most widely applied metric due to its intuitive nature and convenient calculation. It has become a de facto standard to report the Distinct score to compare the performance of different models in terms of response diversity (Liu et al., 2016; Fan et al., 2018; Wu et al., 2021; Zhou et al., 2021). Most previous works follow the initial approach of Li et al. (2016) to calculate the Distinct score, i.e., dividing the number of unique tokens (n-grams) by that of all tokens (n-grams). However, although reported to be effective, we surprisingly find that this naive approach tends to introduce more penalty to longer texts and lead to inaccurate evaluation of the text diversity.

We argue that the scaling factor of *Distinct* requires a comprehensive discussion for two reasons. First, prior research in non-computational linguistics has demonstrated the shortcomings of *Distinct*’s scaling approach (Malvern et al., 2004). We found that early applications of *Distinct* exist in psychological linguistics, where researchers leveraged this metric to assess the language diversity of children with communication disorders (Chotlos, 1944). Their research showed that as a child speaks more words, *Distinct* experiences an adverse decline since each extra word that the child utters adds to the total number of words, yet it would only increase the number of distinct words if the word had not been used before (Malvern et al., 2004; Chotlos, 1944). Second, we also discovered an uncommon decline of this metric on both natural corpus and a designated distribution sampler when the total number of words increases. As illustrated in Figure 1, the original *Distinct* cannot keep a stable value and experiences a sharp decrease with increasing utterance length in both natural and designated distributions. However, as a qualified metric needs to support quantitative comparison among different methods, its value should stay invariant when the distribution of the words appearing is determined. This result is consistent with the findings of psy-

![Figure 1: The original and new Distinct scores against different sample lengths. In the figure, “natural” means that text sets are sampled from a real corpus while “designated” means that the sets are sampled from a designated distribution. See details in Section 2.](http://opus.nlpl.eu/OpenSubtitles2018.php)
chologists, indicating an over-penalty does exist in such a scaling method.

Our contributions are summarized as follows:

1. We investigate the performance of the original Distinct and demonstrate that this metric is not sufficiently fair due to its scaling method. We also highlight the risks of using this metric for evaluating response diversity.

2. We propose an improved version of Distinct (New Distinct) based on the idea that the scaling factor should be the expectation of the number of distinct tokens instead.

3. Human evaluation shows that New Distinct correlates better with human judgments. We further discuss the drawbacks of New Distinct and suggest feasible ways of using this metric in practice.

2 Preliminary Discussion about Original Distinct

To exemplify the shortcoming of origin Distinct, we depicted Distinct score on two kinds of texts at different lengths. One kind of text is sampled from an artificially designated distribution and the other is sampled from a real corpus. In detail, the designated distribution we adopted is \( P(X = k) = \int_0^v \frac{\lambda^ke^{-\lambda}}{e^k!} d\lambda \), where \( v \) is vocabulary size and we simply let it be 30522 (Devlin et al., 2019). The real corpus we adopted is the crawled data from OpenSubtitles\(^1\). For each length, we sampled 2000 sentences as a set and calculated scores of each set.

We found the original Distinct scores decrease sharply with increasing utterance length in both distributions. As shown by the "original-designated" line, with the distribution being determined, lengthier texts will get lower scores than shorter texts. We highlighted this problem because it is extremely simple for models to control the length of texts by using decoding tricks, like adjusting the penalty coefficient (Vijayakumar et al., 2016). It makes a model "easily" beat the other model by such tricks.

3 Improving Original Distinct

3.1 Formula Derivation

The original Distinct score (Li et al., 2016) is measured as \( \text{Distinct} = \frac{N}{C} \), where \( N \) is the number of distinct tokens and \( C \) is the total number of tokens. To improve the original scaling method, we propose that the scaling factor should be the expectation of the number of distinct words in the set of generated responses. Hence, it becomes

\[
\text{NewDistinct} = \frac{N}{\mathbb{E}[\hat{N}]} 
\]

Supposing a set of generated responses \( R \) with size \( S \) to be evaluated, we let \( l_{k,j} \) be the \( j \)th token of \( k \)th response in \( R \) and \( t_k \) be the length of \( k \)th response. The expectation \( \mathbb{E}[\hat{N}] \) for \( \hat{N} \) distinct words to appear in \( R \) would be

\[
\mathbb{E}[\hat{N}] = \sum_j \left( 1 - \prod_k \mathbb{P}(l_{k,t_k} \neq u_j, ..., l_{k,1} \neq u_j) \right),
\]

where \( V \) denotes the vocabulary size, and \( \{u_1, ..., u_V\} \) is the set of all tokens in the vocabulary.

As shown in Equation 2, the calculation requires us to know \( \mathbb{P}(l_{k,t_k} \neq u_j, l_{k,t_k-1} \neq u_j, ..., l_1 \neq u_j) \). Though current models can easily estimate the probability of a word appearing behind given words, it is hard to calculate the probability of each word that never appears in any position of a sequence. Thus, there is may no efficient way to calculate \( \mathbb{P}(l_{k,t_k} \neq u_j, ..., l_{k,1} \neq u_j) \). Besides, different language distributions have different \( \mathbb{P} \), which leads to different expectations and make the metric less general. Thus, we employ the upper bound of response diversity (i.e. a set of generated responses where each token appears with equal probability) to calculate this expectation. We hypothesize that the scaling effect of the upper bound is approximately proportional to that of other sets of generated responses; therefore, it can replace the original scaling factor.

\[
\mathbb{E}[\hat{N}] \leq \mathbb{E}[N^{\text{upper}}],
\]

\[
\mathbb{E}[N^{\text{upper}}] = \sum_j \left( 1 - \prod_k \prod_i \mathbb{P}(l_{k,i} \neq u_j) \right)
\]

\[
= V \left[ 1 - \left( \frac{V - 1}{V} \right)^C \right]
\]
Thus, new \textit{Distinct} score is calculated as:
\begin{equation}
New\text{Distinct} = \frac{N}{V[1 - \left(\frac{1}{V}\right)^C]}
\end{equation}

We have the details of formula derivation, a piece of discussion of the formula’s properties and the determination of vocabulary size in Appendix.

3.2 Experimental Verification

3.2.1 Evaluation Approach

We compared new \textit{Distinct} with the original unigram \textit{Distinct} (Li et al., 2016) by calculating both metrics on the results of ten methods for diversifying dialog generation, reported by Wang et al. (2021). Please see the detailed introduction of the reported methods in Appendix.

As correlation analysis has been widely used to evaluate automatic metrics for language generation (Tao et al., 2018; Sellam et al., 2020), we calculated the Pearson, Spearman, and Kendall correlation coefficients between both scores and human judgments. Pearson’s correlation estimates linear correlation while Spearman’s and Kendall’s correlations estimate monotonic correlation, with Kendall’s correlation being usually more insensitive to abnormal values. We used SciPy\(^2\) for correlation calculation and significance test.

3.2.2 Datasets

Our experiments use two open-domain dialog generation benchmark datasets: DailyDialog (Li et al., 2017), a high-quality dialog dataset collected from daily conversations, and OpenSubtitles\(^3\), which contains dialogs collected from movie subtitles (see Table 1 for more details). We follow the data processing procedures reported by Wang et al. (2021).

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>DailyDialog</td>
<td>65.8K</td>
<td>6.13K</td>
<td>5.80K</td>
</tr>
<tr>
<td>OpenSubtitles</td>
<td>1.14M</td>
<td>20.0K</td>
<td>10.0K</td>
</tr>
</tbody>
</table>

Table 1: Dataset Statistics

3.2.3 Preliminary Observations

Based on the obtained results (check Table 2), it can be observed that NewDistinct has a clear edge over the original Distinct: first, the contrast between diversity of generated responses for different methods is highlighted more effectively by NewDistinct (e.g. though AdaLab gets the highest diversity score using Distinct (3.96), its difference from other methods is not as evident as its NewDistinct score (9.63)); second, in contrast to Distinct, NewDistinct provides a more accurate evaluation of response diversity. For instance, the Distinct scores for CP and UL are both 2.35 while responses generated by UL are found to be more diverse than CP using NewDistinct (5.35 > 5.08). Given that the average length of responses generated by FL is larger than CP, Distinct’s bias towards models that generate shorter sentences becomes evident. These observations are consistent for both datasets.

3.2.4 Correlation Results

We recruited crowdsourcing workers to evaluate the diversity of the selected methods. For each method, we randomly sampled 100 subsets of 15 responses from their set of generated responses. Response sets of all methods, given the same query set, were packaged together as an evaluation set. We asked each crowdsourcing worker to assign a diversity score to every response group in the evaluation set. Each group was evaluated by at least 3 workers. For ensuring the quality of our annotations, we calculated the score of each set as the average of workers’ scores and filtered out workers whose scores had an insufficient correlation with the average (Pearson Correlation < 0.65). We acknowledge that building a scoring standard for annotating language diversity is challenging. Hence, we did not require our workers to give an absolute score for each set. Instead, we asked them to highlight the contrast between different sets by scoring values that linearly reflect the response diversity difference between the sets. For instance, the two sets of scores \{1, 2, 2\} and \{2, 5, 5\} show the same evaluation since the same contrast is shown. We then normalized the scores to the [0-10] range.

Then, we calculated the correlation between the Distinct scores with the crowdsourced values for all the methods. The results are provided in Table 2. The evaluation results indicate that our proposed NewDistinct is more consistent with human judgments for measuring response diversity, as NewDistinct shows the highest correlation with human evaluations among all correlation metrics (Pearson/ Spearman/ Kendall) on both datasets.

\(^2\)https://docs.scipy.org/doc/scipy/reference/stats.html
\(^3\)http://opus.nlpl.eu/OpenSubtitles2018.php
we showed in Figure 1, it is an unfair indicator as it is affected by the sample length. This causes a bias against models which tend to generate longer sentences.

Table 2: Results of automatic and human evaluation on corpus-level diversity methods. Pearson/Spearman/Kendall indicates the Pearson/Spearman/Kendall correlation respectively. The correlation scores marked with †(i.e., p-value<0.1) and ‡(i.e., p-value<0.05) indicate the result significantly correlates with human judgments. The number in parenthesis denotes the standard deviation of response length.

<table>
<thead>
<tr>
<th>Method</th>
<th>DailyDialog</th>
<th>OpenSubtitles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg Length</td>
<td>Distinct</td>
</tr>
<tr>
<td>FL(2017)</td>
<td>9.33</td>
<td>2.38</td>
</tr>
<tr>
<td>NL(2020)</td>
<td>9.99</td>
<td>1.66</td>
</tr>
<tr>
<td>CP(2017)</td>
<td>8.67</td>
<td>2.35</td>
</tr>
<tr>
<td>LS(2016)</td>
<td>8.50</td>
<td>1.48</td>
</tr>
<tr>
<td>D2GPo(2019)</td>
<td>9.15</td>
<td>1.26</td>
</tr>
<tr>
<td>CE(2020)</td>
<td>8.29</td>
<td>1.67</td>
</tr>
<tr>
<td>F2(2020)</td>
<td>8.71</td>
<td>1.40</td>
</tr>
<tr>
<td>UL(2019)</td>
<td>9.93</td>
<td>2.35</td>
</tr>
<tr>
<td>Face(2019)</td>
<td>10.62</td>
<td>1.63</td>
</tr>
<tr>
<td>AdaLab(2021)</td>
<td>11.30</td>
<td>3.96</td>
</tr>
</tbody>
</table>

4 New Distinct in Practice

As new Distinct is based on idealized assumption that does not take language distribution into account, we further discuss this problem and propose a potential practical way of new Distinct in real situations. Before applying new Distinct, it is necessary to explore the relationship between score and text length (Figure 1) and check the performance of Distinct on the training data. To our knowledge, if the training data is from large-scale open-domain sources such as OpenSubtitles and Reddit, Distinct can maintain its value on different lengths. Hence, it can be directly used for evaluating models trained on these datasets. However, we found our experiments on datasets such as Twitter showed a decline in Distinct on lengthier texts. It is probably because of the platform rule of limiting text length under 280, which induces users to say as much information as possible within a shorter length. In this situation, it must be unfair for those methods that tend to generate lengthier texts.

There exist other metrics for evaluating diversity but no one is as widely-used as Distinct (Zhu et al., 2018; Xu et al., 2018). Specifically, Self-BLEU proposed by Zhu et al. (2018) is extremely time-consuming as its computation complexity is $O(n^2)$, where $n$ denoted the size of the test set.

6 Conclusion

In this paper, we proposed an improved variation of the Distinct score, which is a widely-used metric for evaluating response diversity in dialog systems. We provided the theory as well as the methodology behind the formulation of our proposed score (New Distinct). In addition, we conducted experiments on recently proposed dialog generation methods to verify the effectiveness of this metric. The obtained results demonstrated that New Distinct has a higher correlation with human evaluation in comparison with other metrics.

References


A Comparison on More Datasets
To demonstrate the shortcomings of the original Distinct metric, we illustrate original Distinct on 6 datasets: Persona-chat (Zhang et al., 2018), Ubuntu Dialog Corpus (Lowe et al., 2015), DailyDialog, Topic-Chat (Gopalakrishnan et al., 2019), Empathetic Dialogs (Rashkin et al., 2018), Wizard of Wikipedia (Dinan et al., 2018), Reddit (Serban et al., 2015), and Twitter (Ritter et al., 2010) (Figure 1). It can be observed that with an increasing sample length, the original Distinct score tends to follow a linear decline while the proposed metric maintains its consistency.

B Formula Derivation and Property Discussion
\[
E[\hat{N}] = E \left[ \sum_{j} V \left( \sum_{i,k} I_{l_k = u_j} \right) \right]
\]

\[
= \sum_{j} \mathbb{P} \left( \{ I_{l_{k,i} = u_j} \} = 1 \right) \]

\[
= \sum_{j} \left( 1 - \prod_{k} \mathbb{P} (l_k \neq u_j, \ldots, l_1 \neq u_j) \right)
\]
Figure 2: Original scores against different sample lengths. The dotted lines are the actual curves for each score while the lines are slope-intercept graphs of the curves. Each score is calculated based on 10 sets of 2000 randomly sampled responses with the same certain length.

\[ \hat{E}^{\text{Upper}} = V \sum_j \prod_k \prod_i P(l_{k,i} \neq u_j) \] (10)

Thus, our proposed Distinct score is calculated as

\[ \hat{E}^{\text{Upper}} = V \sum_j \prod_k \prod_i P(l_{k,i} \neq u_j) \] (10)

\[ = V[1 - \left(\frac{V - 1}{V}\right)^C], \] (11)

Formula Property 1. NewDistinct increases faster as \( C \) is increasing, but its incremental rate converges to \( \frac{1}{V} \), as shown by its derivative below:

\[ \frac{d\text{NewDistinct}}{dN} = \frac{1}{V[1 - (\frac{V - 1}{V})^C]} \] (13)

\[ \lim_{C \to +\infty} \frac{d\text{NewDistinct}}{dN} = \frac{1}{V} \] (14)

whereas in the original Distinct, we have

\[ \frac{d\text{Distinct}}{dN} = \frac{1}{C} \] (15)

We can see from the original metric that the bigger \( C \) is, the slower the original Distinct increases. It is the reason why this metric is not fair to those models that tend to generate longer sentences.

**Formula Property 2.** NewDistinct converges to \( \frac{N}{V} \) (\( \leq 1 \)) as \( C \) increases.

\[ \lim_{C \to +\infty} \text{NewDistinct} = \lim_{C \to +\infty} \frac{N}{V[1 - \left(\frac{V - 1}{V}\right)^C]} \] (16)

\[ = \frac{N}{V} \leq 1, \] (17)

Details of Human Evaluation

Our created human evaluation interface is provided in Figure 3.

D How to Determine Vocabulary Size

As we discussed the properties of NewDistinct, vocabulary size makes little impact on changing its value when it has reached a large number (usually more than 30000), so it is not necessary to measure an exact value. To compare different methods, it is recommended to use a common vocabulary size, (such as BERT’s 30522) (Devlin et al., 2019). It is also reasonable to calculate the vocabulary size of a
dataset by NLTK tokenizer, when research focuses on a specific dataset. For non-english corpora, we recommend researchers to determine a vocabulary size following Xu et al. (2021).

E  Details of Evaluated Methods

Wang et al. (2021) proposed a novel adaptive label smoothing method for diversified response generation. Their experiments were conducted on the DailyDialog and OpenSubtitles datasets, using 9 recent methods for diverse response generation as their baselines (similar to what we demonstrated in our paper). Wang et al. (2021) used a transformer-based sequence-to-sequence model (Vaswani et al., 2017) as the backbone of their model, and most of their hyper-parameters follow (Cai et al., 2020). In addition, both the encoder and the decoder contain 6 transformer layers with 8 attention heads, and the hidden size is set to 512. BERT’s WordPiece tokenizer (Devlin et al., 2019) and Adam optimizer (Kingma and Ba, 2015) are used for training their models with random initialization and a learning rate of 1e-4.
Evaluating Diversity of Ten Sentence Sets – Manual Evaluation

Task Description

There are ten sentence sets from ten different generative models. You should analyze all the sets and evaluate the diversity of each sentence set by comparing it to others.

You should know:

I. Lexical diversity can be measured by the extent of using various different words in a sentence set. For example, set $A$ ("a d e v s", "g e d h e") is more diverse than set $B$ ("a b c d e", "e d c a b") because set $A$ contains more unique (distinct) words.

III. Though I, please not give your score by counting the number of distinct words for each set because as a sentence is longer, it is harder to increase a distinct word than a shorter sentence. You should evaluate the diversity based on your commonness — whether this sentence at its length is really diverse.

III. You can give each set a score from 1 to 50, where 50 means the highest lexical diversity and 1 means the lowest lexical diversity. For example, you evaluate the lexical diversity of 3 set, $A$, $B$ and $C$, and the result is $A>B>C$. You can give $A$ the highest score (e.g. 40), give $B$ a median score (e.g. 35), and give $C$ the lowest score (e.g. 20).

IV. The absolute score that you give each set is not important, however, the difference between scores should reflect the extent of diversity difference between the sentence sets. For example, if you give $A>5$, $B>9$, $C>10$, that means the difference between $A$ and $B$ (5-9) is much more than that the difference between $B$ and $C$ (0-10). Hence, we can see that $A$ is much less diverse than the others. You can see that the same conclusion could be made if you had scored these three sets as $a>10 b>18 c>20$.

Notes

- Every case is reviewed by more than 5 people. If the rank of the sets that you gave is much different from the results from other workers, we will carefully review your performance again to decide if your task should be accepted. Please ensure that you take it seriously.

Assignment: evaluate the diversity of each sentence set by comparing it to others.

Set 1:
1. There’s no way to nail them.
2. I’ll be back in a minute.
3. Though, he replied, “I’m gon na be able to make a wish.”
4. We’re going to go to the forest.
5. I don’t care.
6. We got a little problem.
7. I’ll be there.
8. I’ll ride him.
9. How could it be?
10. I’m not afraid.
11. I mean, I was trained to get him out of prison.
12. I’m gon na get you out of here.
13. I’m here to see you.
14. I don’t know.
15. I got to get to the embassy.
On a scale of 1-50, how much lexical diversity score do you think this set gets? ____________

Set 2:
1. The judges will be here by the next day.
2. I’ll just go to the movies.
3. So, she’d be happy to be able to communicate with her.
4. We have to go.
5. I’ll give you $50.
6. We got a problem.
7. We’re all right.
8. I bet he will.
9. How could he have been involved with the computer?
10. I’m not sure.
11. But I was still alive.
12. I’m not finished.
13. I’m here to see you.
14. She was at the scene.
15. I’ll take care of it.
On a scale of 1-50, how much lexical diversity score do you think this set gets? ____________

Set 3:
1. And they will show up to you, and you will be back in a few minutes.
2. I’m not sure.
3. The word is kateina, to have seen the kates.
4. We have to go to war.
5. I’ll take it.
6. We’re in the same area.
7. I’m gon na have some fun.
8. I’m sure he’ll have a horse.
9. What kind of flies?
10. I’m not a bad person.
11. I thank you, Mr. Bond.
12. I’m not sure I’m gon na do it.
13. I’m here to see you.
14. They’re not in charge of this investigation.
15. I’m going to kill you all.
On a scale of 1-50, how much lexical diversity score do you think this set gets? ____________

Figure 3: Interface of Human Evaluation