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-

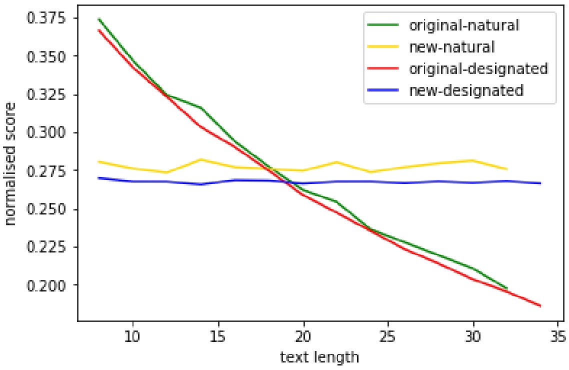


Figure 1: The original and new *Distinct* scores against different sample lengths. In the figure, "natural" means that text sets are sampled from a real corpus while "designated" means that the sets are sampled from a designated distribution. See details in Section 2.

[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 $\mathbb{P}\left(X=k\right)=\int_0^v \frac{\lambda^k e^{-\lambda}}{vk!}d\lambda$, 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 while it obviously not suitable to draw the conclusion that a model performs better than the other in diversity. The same phenomenon can be observed on the real corpus (see "original-natural" line in Figure 1). As language distribution is more complex than what we are able to formulate, we depicted the performance of the original *Distinct* on 6 famous datasets in **Appendix**. Many cases indicate that the original *Distinct* is really unreasonable to be a fair metric for evaluating diversity.

3 Improving Original Distinct

3.1 Formula Derivation

The original *Distinct* score (Li et al., 2016) is measured as $Distinct=\frac{N}{C}$, 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

$$NewDistinct=\frac{N}{\mathbb{E}\left[\hat{N}\right]} \qquad (1)$$

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

$$\mathbb{E}\left[\hat{N}\right]=\sum_j^V(1-\prod_k^S\mathbb{P}\left(l_{k,t_k}\neq u_j,...,l_{k,1}\neq u_j\right)), \qquad (2)$$

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 $\mathbb{P}\left(l_{t_k}\neq u_j,l_{t_k-1}\neq u_j,...,l_1\neq u_j\right)$. 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 $\mathbb{P}\left(l_{k,t}\neq u_j,...,l_{k,1}\neq u_j\right)$). 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.

$$\mathbb{E}\left[\hat{N}\right]\underset{\sim}{\propto}\mathbb{E}\left[\hat{N_{upper}}\right], \qquad (3)$$

$$\begin{aligned}\mathbb{E}\left[\hat{N_{upper}}\right]&=\sum_j^V(1-\prod_k^S\prod_i^{t_k}\mathbb{P}\left(l_{k,i}\neq u_j\right)) &(4)\\ &=V[1-(\frac{V-1}{V})^C] &(5)\end{aligned}$$

[1] http://opus.nlpl.eu/OpenSubtitles2018.php

Thus, new *Distinct* score is calculated as:

$$NewDistinct = \frac{N}{V[1 - (\frac{V-1}{V})^{C}]} \tag{6}$$

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 *Distinct* with the original unigram *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).

	Train	Val	Test
DailyDialog	65.8K	6.13K	5.80K
OpenSubtitles	1.14M	20.0K	10.0K

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/ Spearson/ Kendall) on both datasets.

[2] https://docs.scipy.org/doc/scipy/reference/stats.html

[3] http://opus.nlpl.eu/OpenSubtitles2018.php

[3] See Appendix for more details on the human evaluation interface

Method	DailyDialog				OpenSubtitles			
	Avg Length	Distinct	Ours	Human	Avg Length	Distinct	Ours	Human
FL(2017)	9.33	2.38	5.09	5.18	8.56	3.19	9.51	4.91
NL(2020)	9.99	1.66	3.70	4.54	8.40	3.24	9.52	5.02
CP(2017)	8.67	2.35	4.80	5.08	8.74	3.11	9.44	5.20
LS(2016)	8.50	1.48	2.98	5.28	9.04	2.77	8.64	5.04
D2GPo(2019)	9.15	1.26	2.65	4.92	8.77	2.07	6.32	4.89
CE(2020)	8.29	1.67	3.31	4.14	9.21	2.55	8.08	4.95
F^2(2020)	8.71	1.40	2.87	4.88	8.60	2.89	8.67	4.52
UL(2019)	9.93	2.35	5.23	5.35	8.09	2.84	8.10	5.00
Face(2019)	10.62	1.63	3.79	5.26	9.11	3.31	10.41	5.31
AdaLab(2021)	11.30	3.96	9.63	5.92	8.12	4.78	13.68	5.32
Pearson	-	**0.67‡**	**0.70‡**	**1.00**	-	**0.56†**	**0.60†**	**1.00**
Spearman	-	**0.42†**	**0.62†**	**1.00**	-	**0.62†**	**0.65‡**	**1.00**
Kendall	-	**0.27**	**0.47†**	**1.00**	-	**0.51‡**	**0.56‡**	**1.00**

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.

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.

5 Related Work

Li et al. (2016) proposed *Distinct*, calculated as the number of distinct tokens divided by the total number of tokens. To our knowledge, this is the most widely-used automatic metric for evaluating response generation diversity. However, as 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.

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

Hengyi Cai, Hongshen Chen, Yonghao Song, Cheng Zhang, Xiaofang Zhao, and Dawei Yin. 2020. Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6334–6343, Online. Association for Computational Linguistics.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 7893–7905, Online. Association for Computational Linguistics.

Byung-Ju Choi, Jimin Hong, David Park, and Sang Wan Lee. 2020. F^2-softmax: Diversifying neural text generation via frequency factorized softmax. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9167–9182, Online. Association for Computational Linguistics.

John W. Chotlos. 1944. Iv. a statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75–111.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinglang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, Dilek Hakkani-Tür, and Amazon Alexa AI. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. In *INTERSPEECH*, pages 1891–1895.

Tianxing He and James Glass. 2020. Negative training for neural dialogue response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2044–2058, Online. Association for Computational Linguistics.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference*, pages 2879–2885.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*, pages 110–119.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2019. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael 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 *EMNLP*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical diversity and language development*. Springer.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Alan Ritter, Colin Cherry, and William B Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021. Diversifying dialog generation via adaptive label smoothing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3507–3520.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. 2021. Vocabulary learning via optimal transport for neural machine translation. 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 7361–7373, Online. Association for Computational Linguistics.

Zhen Xu, Nan Jiang, Bingquan Liu, Wenge Rong, Bowen Wu, Baoxun Wang, Zhuoran Wang, and Xiaolong Wang. 2018. LSDSCC: a Large Scale Domain-Specific Conversational Corpus for Response Generation with Diversity Oriented Evaluation Metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2070–2080, Stroudsburg, PA, USA. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664.

Hao Zhou, Pei Ke, Zheng Zhang, Yuxian Gu, Yinhe Zheng, Chujie Zheng, Yida Wang, Chen Henry Wu, Hao Sun, Xiaocong Yang, et al. 2021. Eva: An open-domain chinese dialogue system with large-scale generative pre-training. *arXiv preprint arXiv:2108.01547*.

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

A Comparison on More Datasets

To demonstrate the shortcomings of the original Distint 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

$$\mathbb{E}\left[\hat{N}\right] = \mathbb{E}\left[\sum_{j}^{V} \bigvee_{i,k}^{i=t_k,k=S} \mathbb{1}_{l_{k,i}=u_j}\right] \tag{7}$$

$$= \sum_{j}^{V} \mathbb{P}\left(\{\bigvee_{i,k}^{i=t_k,k=S} \mathbb{1}_{l_{k,i}=u_j}\} = 1\right) \tag{8}$$

$$= \sum_{j}^{V} (1 - \prod_{k}^{S} \mathbb{P}\left(l_{t_k} \neq u_j, ..., l_1 \neq u_j\right)) \tag{9}$$

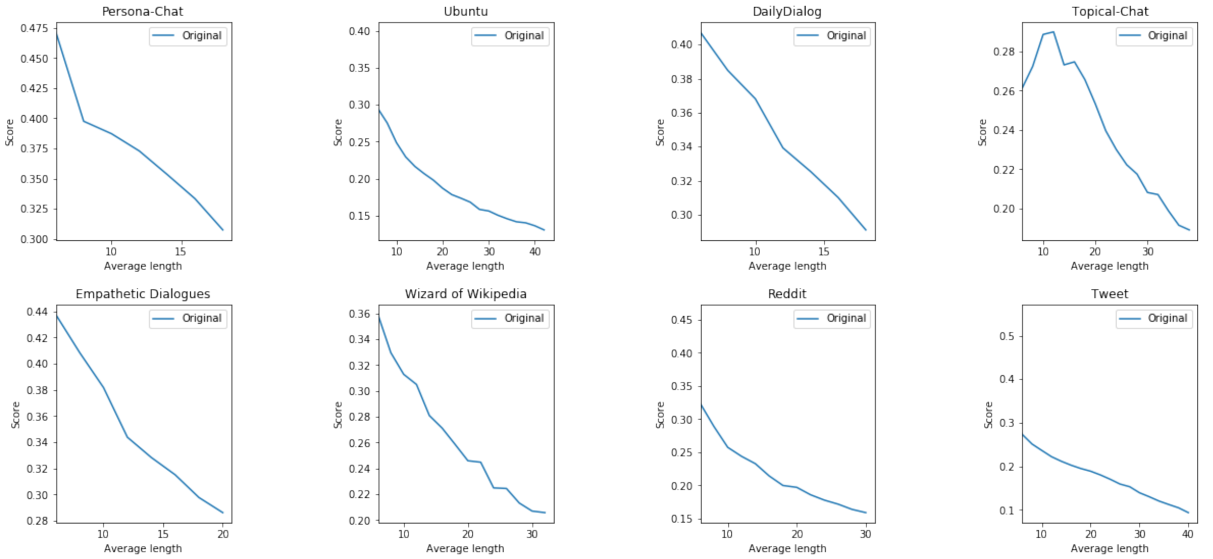


Figure 2: Original scores against different sample lengths. The dotted lines are the actual curves for each score while the lines are slope-intercept graphs of the curves. Each score is calculated based on 10 sets of 2000 randomly sampled responses with the same certain length.

$$\mathbb{E}\left[N_{\hat{upper}}\right] = \sum_{j}^{V}(1 - \prod_{k}^{S}\prod_{i}^{t_k}\mathbb{P}\left(l_{k,i} \neq u_j\right)) \tag{10}$$

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

Thus, our proposed Distinct score is calculated as

$$NewDistinct = \frac{N}{V[1 - (\frac{V-1}{V})^C]} \tag{12}$$

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:

$$\frac{\mathrm{d}NewDistinct}{\mathrm{d}N} = \frac{1}{V[1 - (\frac{V-1}{V})^C]} \tag{13}$$

$$\lim_{C\to+\infty}\frac{\mathrm{d}NewDistinct}{\mathrm{d}N} = \frac{1}{V} \tag{14}$$

whereas in the original Distinct, we have

$$\frac{\mathrm{d}Distinct}{\mathrm{d}N} = \frac{1}{C} \tag{15}$$

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 $\frac{N}{V}$ (≤ 1) as C increases.

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

$$= \frac{N}{V} <= 1, \tag{17}$$

where $\frac{N}{V[1-(\frac{V-1}{V})^C]} \in [0, +\infty]$. Theoretically, NewDistinct can have values larger than 1 (e.g. when $N = V$), which is an extremely rare case in practice: as we utilized the upper bound for measuring the expectation, it is exceptionally hard for N to obtain an equal value to or an even greater value than $\mathbf{E}(N_{\hat{upper}})$.

C Details of Human Evaluation

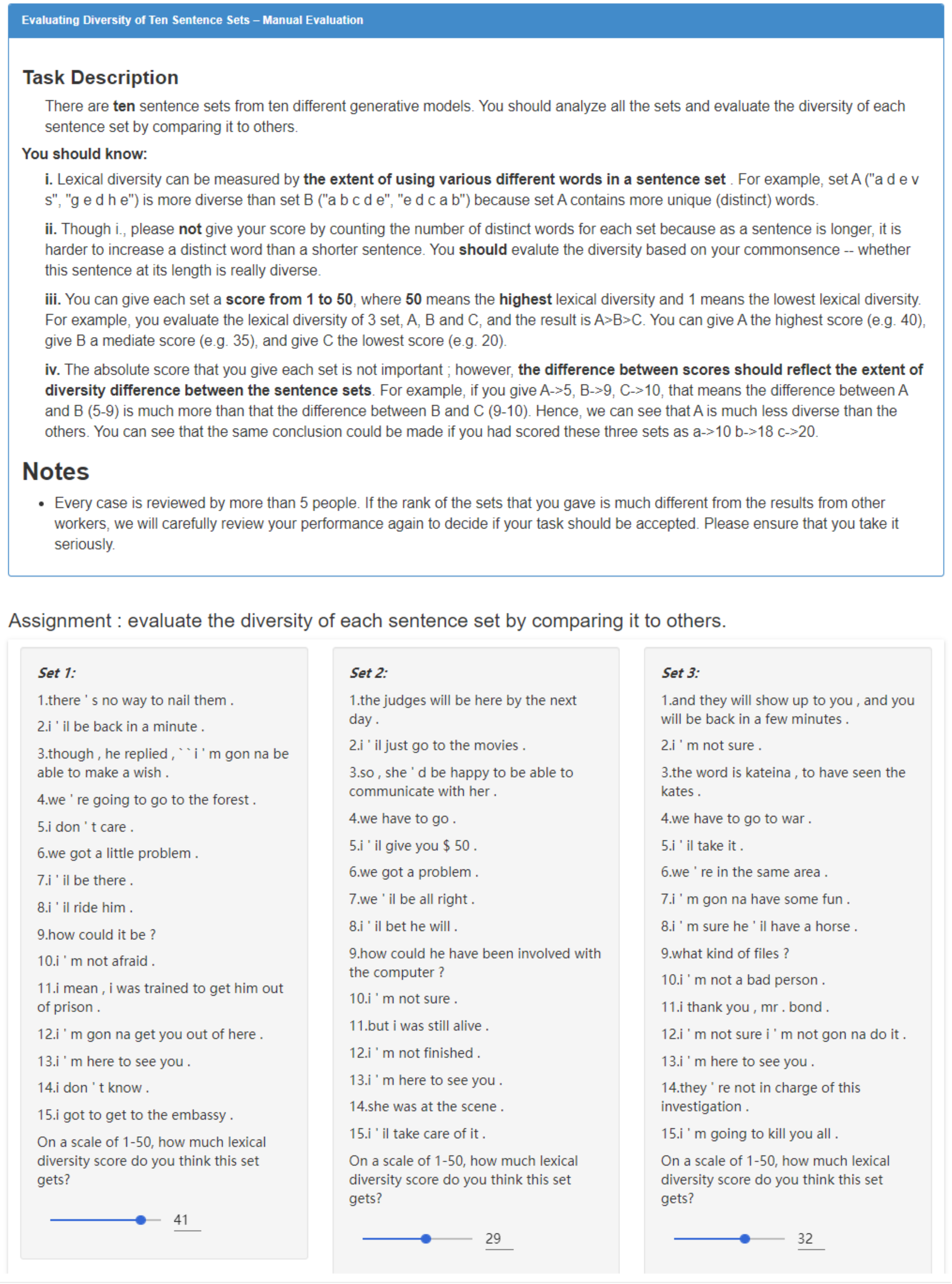
Our created human evaluation interface is provided in Figure 3.

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.

ii. 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** evalute the diversity based on your commonsence -- 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 mediate 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 (9-10). 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 ' il 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 ' il be there .

8.i ' il 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?

41

Set 2:

1.the judges will be here by the next day .

2.i ' il 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 ' il give you $ 50 .

6.we got a problem .

7.we ' il be all right .

8.i ' il 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 ' il take care of it .

On a scale of 1-50, how much lexical diversity score do you think this set gets?

29

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 ' il take it .

6.we ' re in the same area .

7.i ' m gon na have some fun .

8.i ' m sure he ' il have a horse .

9.what kind of files ?

10.i ' m not a bad person .

11.i thank you , mr . bond .

12.i ' m not sure i ' m not 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?

32

Figure 3: Interface of Human Evaluation