Investigating Relationships between Accuracy and Diversity in Multi-Reference Text Generation

Weike Fang
wfang@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

Meng Jiang
mjiang2@nd.edu
University of Notre Dame
Notre Dame, Indiana, USA

ABSTRACT

In text generation, we aim to produce outputs that are not only correct but also diverse in terms of content, use of words, and meaning. The ability to generate accurate and diverse text is crucial in conversation systems, story generation, machine translation, paraphrasing, commonsense reasoning, etc. To efficiently evaluate the generated text, researchers have extensively studied automatic evaluation metrics to substitute expensive, slow human evaluation. Existing metrics include *n*-gram-based metrics and neural-based metrics. The former perform well on measuring form or lexical quality and diversity while the latter excel at detecting semantic quality and diversity, both showing good correlation with human judgments. In this work, we observe that the trade-off between semantic quality and diversity occurs in the output of models trained for multi-reference text generation, making it hard to find the optimal model by looking at quality and diversity metrics separately. We propose a human study framework and provide methods to generate experiment data for researchers to design or evaluate new metrics in the future.

ACM Reference Format:
Weike Fang and Meng Jiang. 2022. Investigating Relationships between Accuracy and Diversity in Multi-Reference Text Generation. In *Proceedings of KDD Undergraduate Consortium (KDD-UC)*. ACM, New York, NY, USA, 7 pages.

1 INTRODUCTION

Human languages, as mediums and means of communication, were initially created so that people can convey thoughts and ideas accurately. In our daily life, there are many scenarios where we also need diversity in written or oral communication. For example, teachers may wish to see diversified stories and ideas under the same essay topic. Audience may wish to hear different answers to the same question from different celebrities on a talk show. Journalists may strive to come up with a wide variety of questions on the same incident for their interviewees or propose unique titles for a report article so that it can appeal more readers.

Producing diverse content requires one person for an extended period or multiple persons to generate several stories under the same topic. It is challenging for both humans and machines. Text generation models are desired to produce outputs that are not only correct, but also diverse [23]. In the literature, the term "diversity" is often referred to as the ability of a generation model to create a set of outputs that are valid and also vary in terms of content, use of words, and meaning [9, 27]. Such a research problem is often referred as *one-to-many generation* [4, 21, 22, 27]. For example, in a conversation system, an engaging generation model should be capable of outputting grammatical, coherent responses that are interesting and diverse, avoiding trivial, commonplace responses [13]. In news article queries, diversifying clickable queries will expose users to a wide variety of article contents, boosting user experience and attracting more users [22]. Other tasks concerning diversity include paraphrase generation [7] and machine translation [21].

Human evaluation is often a good indicator of the quality [20] including diversity of text generated by a system. However, human evaluation is a high-latency and expensive process that does not fit in model development pipeline [20]. Thus, there has been extensive studies about *automatic evaluation metrics* in NLG that approximate human evaluation while being computationally cheap.

The first generation of metrics are *n*-gram-based metrics. Quality *n*-gram metrics include BLEU [18], ROUGE [15], and METEOR [2]. They measure the lexical similarity between sentences relying on handwritten rules such as *n*-gram overlap. Diversity *n*-gram metrics include distinct *n*-gram (distinct-n) [13], Entropy (Ent-n) [30], and Self-BLEU [31]. Since these metrics only concern lexical variation, they cannot appropriately reward semantic or syntactic variations [20]. They perform well on measuring form or lexical diversity but poorly on content or semantic diversity [23].

Researchers address such problem by incorporating learned elements into evaluation metrics. There are metrics that are fully trained to correlate with human evaluation such as BEER, RUSE, and ESIM. Some other metrics such as YiSi and BERTScore [29] combine trained contextual embedding and handcrafted token alignment logic. Sentence-BERT (*sent*-BERT or *SBERT*) [19], for example, is a modification of the pre-trained BERT by using siamese and triplet network structures to construct semantically meaningful sentence embeddings, which can be compared using cosine similarity or Euclidean distance. Such neural metrics are found to outperform *n*-gram metrics at detecting accuracy and diversity [23, 29].

Recently, neural metrics have achieved high correlation with human evaluation in terms of quality [19, 20, 29] and content diversity [23]. In generation tasks, we often hope that the system outputs can be both *accurate* with respect to reference(s) and *diverse* against other generated hypotheses, especially when there are multiple acceptable answers. However, machine often fail to generate accurate and diverse responses comparable to human-crafted references [8, 23]. Moreover, it is observed in our experiments that there is a tradeoff between accuracy and diversity. But there is

KDD-UC, August 14-18, 2022, Washington DC, USA

ACM ISBN 978-1-4503-9999-9/18/06…$15.00

Input: *Lisa played Jazz music for Moses of the Old Testament yesterday.*
Reference 1 ($R^{(1)}$): *Moses is not alive.*
Hypothesis 1 ($H^{(1)}$): *Moses is not alive.*
Reference 2 ($R^{(2)}$): *Moses has been dead for a long time and could not have heard a concert yesterday.*

Options of Hypothesis 2		Accuracy (↑)			Diversity (↑)	Tradeoff observed
		$(H_i^{(2)}, R^{(1)})$	$(H_i^{(2)}, R^{(2)})$	max	$(H_i^{(2)}, H^{(1)})$	
$H_a^{(2)}$:	*Moses is not alive.*	0.94	0.60	0.94	1.06	Too low diversity
$H_b^{(2)}$:	*Moses died years ago.*	0.69	0.62	**0.69**	1.45	Q1: which has better semantic quality?
$H_c^{(2)}$:	*Moses of the Old Testament could not hear music.*	0.67	0.62	0.67	**1.49**	Too low accuracy
$H_d^{(2)}$:	*Lisa did not know how to play jazz music.*	0.26	0.32	0.32	3.82	Q2: how to optimize?
$H_e^{(2)}$:	$[R^{(2)}]$	0.63	1	1	1.59	

Human judgement needed.

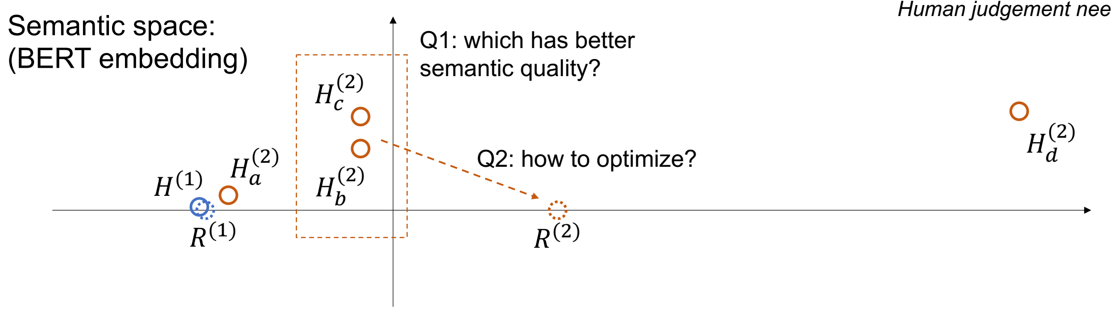


Figure 1: An example of the task of commonsense explanation generation where we observe the tradeoff between accuracy and diversity. *Top*: Table of input, reference, and generated sentences including the accuracy and diversity scores for each hypothesis using sent-BERT [19]. *Bottom*: Sentence embedding based on accuracy and diversity values on the right. In this case, we have two diverse references $R^{(1)}$ and $R^{(2)}$. Suppose Hypothesis 1 $H^{(1)}$ is equivalent to $R^{(1)}$. Multiple options of $H^{(2)}$ are generated. $H_a^{(2)}$ has the highest accuracy but very low diversity. $H_d^{(2)}$ is a wrong response but gives the highest diversity. $H_b^{(2)}$ and $H_c^{(2)}$ exhibit the tradeoff, making it hard for existing metrics and even human to decide which one is better.

no standard method to find the optimal point when the tradeoff occur. A typical scenario would be model selection when training a text generation model. When multiple models are evaluated on the validation set and tradeoff between accuracy and diversity is observed, how do we decide which has the best performance?

Figure 1 is an example in the commonsense explanation generation dataset (ComVE). Given a counterfactual statement, the task aims to generate multiple reasons or explanations about why this statement does not make sense [25]. We use cosine similarity as accuracy and the inverse of cosine similarity as diversity. Accuracy of a hypothesis given two references is the *higher* of the two cosine similarity values with respect to each reference.

Previous works [1, 3] have shown that accuracy (BLEU) and lexical diversity (Self-BLEU) are negatively correlated in single-reference tasks. However, such a trade-off should be expected: generated responses of high accuracy should be clustered close to the single reference while highly diverse responses tend to be found farther from the single reference. In this work, we conduct a systematic experiment on multi-reference tasks to verify the asynchronous behavior of accuracy and diversity from the semantic perspective. Finally, we propose a human study framework to cope with the problem of the trade-off between accuracy and diversity and prepare experiment data from a multi-reference dataset. We will use the data to conduct human studies. Our goal is to learn how human make decisions when accuracy and diversity metrics disagree. By measuring human judgments on the latent space (such as decision boundary), we hope to develop metrics that better align with human evaluation.

2 RELATED WORK AND PRELIMINARIES

2.1 Diversity-Promoted Text Generation

There are many applications for diverse generation (i.e., producing several outputs given a source sequence) such as dialogue system [6, 13], machine translation [21], and story generation [27]. For example, as opposed to the common intuition that translation is a one-to-one mapping, there are actually many plausible translations of the same input differing in style, grammar, or vocabulary [11].

There has been much effort on enhancing diversity in text generation with different approaches. Sampling-based decoding methods are simple yet effective ways to diversify generated text. For example, nucleus sampling [8] samples from the dynamic nucleus of tokens containing the vast majority of the probability mass rather than sampling directly from the probabilities predicted by the model. Another way to improve diversity involve introducing

Table 1: Symbols and their description.

Symbol	Description
m_r, m_h	Number of references/hypotheses
$R^{(i)}$	The i-th reference
$\mathcal{R}$	The set of references: $\mathcal{R} = \{R^{(1)}, ..., R^{(m_r)}\}$
$H^{(i)}$	The i-th hypothesis
$\mathcal{H}$	The set of hypotheses: $\mathcal{H} = \{H^{(1)}, ..., H^{(m_h)}\}$
k	Length of sentence (i.e., number of tokens)
X	An input $X = \{x_1, ..., x_k\}$ of k tokens
n	Size of embedding space

noise or modifying latent vectors [7, 11], such as incorporating variational autoencoder (VAE) to generate diverse text [7, 26].

2.2 Accuracy and Diversity Evaluation

Table 1 presents the symbols (and their descriptions) that we use throughout the paper.

In text generation, two aspects are considered when evaluating generated outputs: *accuracy* and *diversity*. *Accuracy* tests the correctness of the responses with respect to the input context and references, and *diversity* measures the form (lexical) and content (semantic) heterogeneity of the generated responses given the same input. These evaluation metrics have been extensively used in existing literature [4, 24, 27, 31]

2.2.1 n-gram Accuracy Metrics. Popular n-gram metrics, such as BLEU [18], ROUGE [15], and METEOR [2], measure accuracy based on n-gram overlap between the generated hypothesis and the target(s). In a multi-reference setting, given a set of m references $\{R^{(1)}, \ldots, R^{(m_r)}\}$ and a generated hypothesis H, the accuracy is the best accuracy achieved with a reference in the set. Concretely:

$$\text{Accuracy}(H) = \max_{1 \leq i \leq m} \text{Accuracy}(H, R^{(i)})$$

2.2.2 n-gram Diversity Metrics. A popular diversity metric is *distinct n-gram*, which measures the proportion of distinct n-grams out of the total number of n-grams in the corpus [13]. Entropy-n [30] tests lexical diversity by measuring how evenly distributed are the n-grams in a given sentence, with word frequency taken into account. For these two metrics, higher values indicate better diversity.

Another line of n-gram metrics measure pairwise diversity. Self-BLEU [31] computes the BLEU score of a generated sequence with respect to another generated sequence, reflecting the level of similarity inside the set of generated sentences. Concretely, if $\mathcal{H} = \{H^{(1)}, H^{(2)}, \ldots, H^{(m_h)}\}$ are generated given the source sequence, then Self-BLEU computes the average BLEU among all pairwise combinations of $\mathcal{H}$. Low average Self-BLEU implies low similarity (or high diversity) between generated sentences.

2.2.3 Neural Quality Metrics. A new line of metrics aim to use language models such as BERT or generation models such as BART. Some embed generated sequences to a latent space and then evaluate them using geometric features in the space such as cosine similarity or Euclidean distance. Some evaluate generated sentences with the generation probability distribution and calculate the generation probability from different perspectives [28].

2.2.4 Diversity as Similarity Reduction. Tevet et al. proposed a method to construct a diversity metric from any pair-wise similarity metric on sentences [23]. Given a symmetric similarity metric $sim(H^{(i)}, H^{(j)})$ that measure the similarity of a pair of sentences $(H^{(i)}, H^{(j)})$, we can construct a diversity metric $div(\cdot)$ as the negation of the mean similarity score across all pairs of $\mathcal{H}$.

$$div(\mathcal{H}) = -\frac{1}{\binom{|\mathcal{H}|}{2}} \sum_{H^{(i)}, H^{(j)} \in \mathcal{H}, i<j} sim(H^{(i)}, H^{(j)})$$

2.2.5 Neural Diversity metrics. By applying the reduction process above, we can derive corresponding diversity metrics for each neural quality metrics. For example, the Self-BERT metric used in [5, 16], converted from BERTScore, measure the semantic similarity of generated text with itself. The negative sign is omitted to align with self-similarity metrics like Self-BLEU. Lower Self-BERTScore means higher diversity as the generated text are less similar.

$$\text{self-BERT}(\mathcal{H}) = \frac{1}{\binom{|\mathcal{H}|}{2}} \sum_{H^{(i)}, H^{(j)} \in \mathcal{H}, i>j} \text{BERT-Score}(H^{(i)}, H^{(j)})$$

3 VALIDATING TRADEOFF WITH MULTIPLE REFERENCES

Existing work has noticed the tradeoff between accuracy and diversity under n-gram metrics such as BLEU and self-BLEU in single-reference text generation tasks [1, 3]. In those tasks, the tradeoff between accuracy and diversity should be expected since high-quality outputs should converge to the single reference. To the best of our knowledge, there was no work that experiments on multi-reference text generation tasks. Past research recognize that BLEU is not always a good proxy of sample accuracy [3], neither with Self-BLEU for diversity. With the emergence of neural metrics that align better with human semantic evaluation, we conduct a systematic experiment to see if the tradeoff occurs.

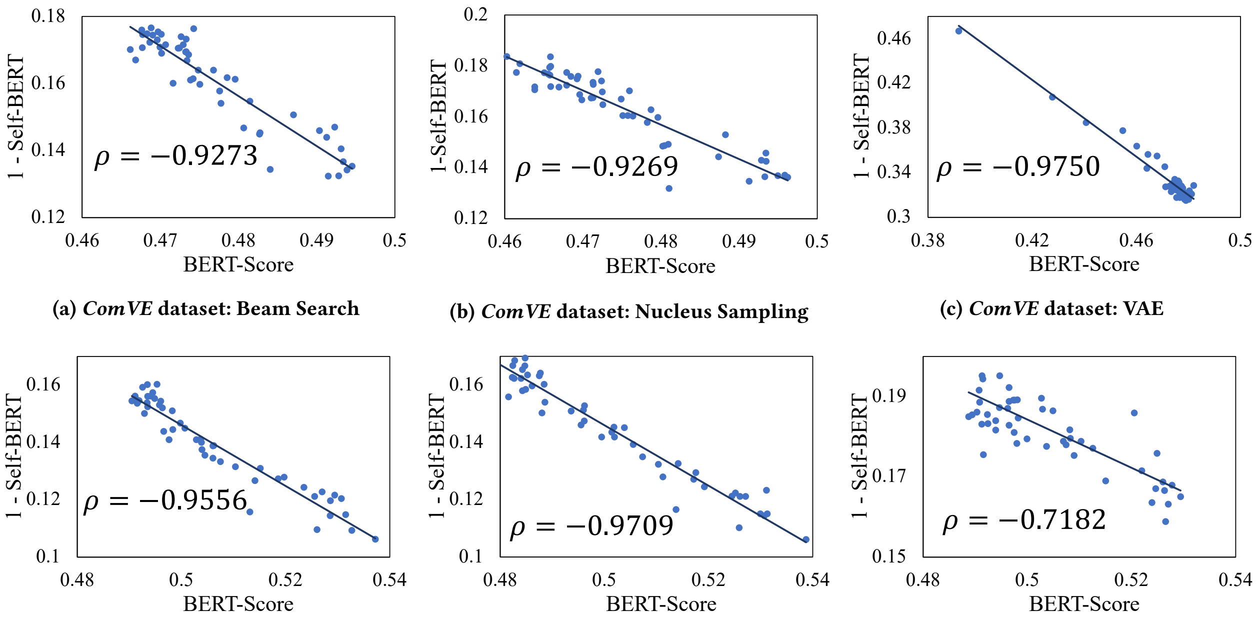
3.1 Text Generation Tasks

We consider two text generation tasks with multi-reference datasets.

3.1.1 Commonsense Explanation Generation. Given a counterfactual statement, the task aims to generate the reason and explanation about why this statement does not make sense [25]. We use the ComVE dataset from SemEval-2020 Task 4 subtask C [25], containing 10,000 / 997 / 1,000 examples for training / dev / test set respectively. Each example in ComVE has three references.

3.1.2 Commonsense Reasoning Generation. CommonGen is a constrained generation task that requires machines to generate a sentence describing common, day-to-day scene using concepts from a given concept set [14]. The CommonGen data contain 32,651 / 993 / 1,497 examples for train / dev / test set respectively. The concept set of each example has three to five concepts.

Figure 2: Diversity vs. accuracy results of models with three decoding strategies on two datasets. Each point on the plot represents the performance of a model on the dev sets. Diversity is measured by 1−Self-BERTScore. Accuracy is measured by BERTScore. Higher values on the two axes represent better diversity and accuracy, respectively.



(a) *ComVE* dataset: Beam Search

(b) *ComVE* dataset: Nucleus Sampling

(c) *ComVE* dataset: VAE

(d) *CommonGen* dataset: Beam Search

(e) *CommonGen* dataset: Nucleus Sampling

(f) *CommonGen* dataset: VAE

3.2 Model Development Details

We use BART, a pre-trained Transformer model that has exhibited great performance on a variety of text generation tasks [12]. Specifically, we utilized BART-base pre-trained weights to initialize the models. They consist of encoders and decoders with 6 layers and 12 attention heads with hidden size of 768. For fine-tuning, we use Adam optimizer with learning rate of 3e-5 that warms up over the first 10k steps ($\beta_1 = 0.9, \beta_2 = 0.999$). For the training process, our models are trained on Tesla V100 GPU with a 4-card 32GB memory.

3.3 Decoding Strategies

We investigate three decoding strategies that enable a text generation model to create different outputs for the same input. With two text generation tasks/datasets and three strategies, we wish that our investigation leads to reliable observations.

3.3.1 Beam Search. Beam search is a search algorithm that stores B highest-scored partial solutions at each time step (B is the *beam width*). At the time step t, beam search looks at all possible token extension of existing beams and retains the B beam tokens [24].

3.3.2 Nucleus Sampling. Nucleus sampling truncates the unreliable tail of the probability distribution and sample from the dynamic nucleus of tokens containing the vast majority of the probability mass to avoid text degeneration [8].

3.3.3 VAE. Variational auto-encoder (VAE) [10] is a generative latent variable model. Recently, VAE has been used to build generative frameworks to generate diverse responses by sampling latent variables from an approximate posterior distribution [7, 26].

3.4 Observations

Figure 2 shows the relationship between between BERT-Score (accuracy) and 1−Self-BERT (diversity) using BART models with beam search, nucleus sampling, and VAE decoding strategies. We have a few interesting observations.

First, we observe that the diversity score is **consistently negatively correlated** with the accuracy score on the multi-reference text generation tasks. There is no optimal model that outperforms others in terms of both accuracy and diversity.

Second, BART models with the **VAE decoding strategy** produce much more diverse responses on the text generation tasks at the expense of a small amount of accuracy. Beam search and nucleus sampling create similar levels of semantic accuracy, but nucleus sampling gives models the ability to achieve higher diversity.

Third, we find that the accuracy does **not synchronize** with diversity with respect to **training epochs**. Specifically, the epoch that achieves the maximal accuracy does not achieve the best diversity; and the epoch that achieves the best diversity does not achieve the best accuracy. Due to the asynchronous behavior, it is hard to find an absolute optimum considering both accuracy and diversity.

Discussion. From the fact that both metrics align well with human evaluation [5, 16, 29], we assume that human evaluation of accuracy and diversity separately would exhibit a similar pattern. However, humans are able to detect when a set of responses is too diverse to be acceptable (i.e., unacceptably wrong), while unfortunately this cannot be reflected by automatic metrics.

Therefore, we are interested in investigating how humans make decisions on the semantic quality of machine-generated text, when there is a tradeoff between accuracy and diversity. We aim to understand how human makes decisions in this case so that we can simulate such behaviors to a unified scoring or measurement across training epochs or develop a new metric that aligns better with human judgment.

4 HANDLING THE TRADEOFF: PROPOSED APPROACH USING HUMAN EVALUATION

We propose an approach to collect human evaluation data and analyze them to handle the tradeoff between accuracy and diversity. This approach is designed towards any multi-reference text generation task and any embedding space.

Given a set of references $\mathcal{R} = (R^{(1)}, R^{(2)}, \dots, R^{(m)})$ $(m \in \{2, 3\})$ and two sets of hypotheses $\mathcal{H} = (H^{(1)}, H^{(2)}, \dots, H^{(m)})$ and $\mathcal{H}' = (H^{(1)'}, H^{(2)'}, \dots, H^{(m)'})$, humans are able to decide which set of hypotheses is better if one set outperforms the other on both accuracy and diversity. Then a key question that we wish to answer is: if one set of hypotheses have better accuracy and the other is more diverse, how can we decide which set is better overall?

4.1 Assumptions

In our analysis, we have a few assumptions about semantic diversity. First, as long as machine responses is as diverse as human generated text, there is no need to aim for higher diversity. The diversity of human hand-written text is indicated by that of ground-truth references in multi-reference tasks. In most cases, ground truths are highly diverse. Besides, given the first hypothesis sentence $H^{(1)}$, if we further generate two options of the second hypothesis $H_a^{(2)}$ and $H_b^{(2)}$ of the same accuracy score, then we prefer the one that is more diverse compared to $H^{(1)}$. The same holds when we desire three or more generations. Existing diversity evaluation metrics like Self-BERTScore don't take into account accuracy. Higher diversity may not always mean better quality. Random output may give a very high diversity but its accuracy is not acceptable compared with the references.

4.2 Experiment Setup

In our experiment design, we tackle the problem by controlling all but one hypotheses in a set to be evaluated by a human annotator, instead of comparing two completely different sets of hypotheses. Specifically, we design our experiment as below: Given $\mathcal{R} = \{R^{(1)}, R^{(2)}\}$ and $H^{(1)}$ that is very close to $R^{(1)}$, we generate a set of $H^{(2)}$ and let humans decide which one yields the best combination with $H^{(1)}$.

4.3 Data Preparation

We aim to find a set of $H^{(2)}$ for each data instance for human to evaluate their quality. In each set, any two hypotheses are expected to exhibit a tradeoff in accuracy and diversity, where human annotators are asked to rate them.

4.3.1 Overview. Figure 4 presents how we find multiple options of $H^{(2)}$. Suppose we are given $R^{(1)}$ and $R^{(2)}$. l is a hyperplane on which the points have equal distance to $R^{(1)}$ and $R^{(2)}$. We call it *the equidistant hyperplane* (of $n-1$ dimensions in an n-dimensional embedding space). $\{l_i\}$ are hyperplanes that are parallel to the equidistant hyperplane l. Hypothesis points on a particular l_i exhibit a tradeoff between accuracy and diversity. $H_c^{(2)}$ gives high diversity (far from $H^{(1)}$) but low accuracy (close to neither reference). In comparison, $H_a^{(2)}$ has higher accuracy and lower diversity since it is closer to $R^{(2)}$ and $H^{(1)}$.

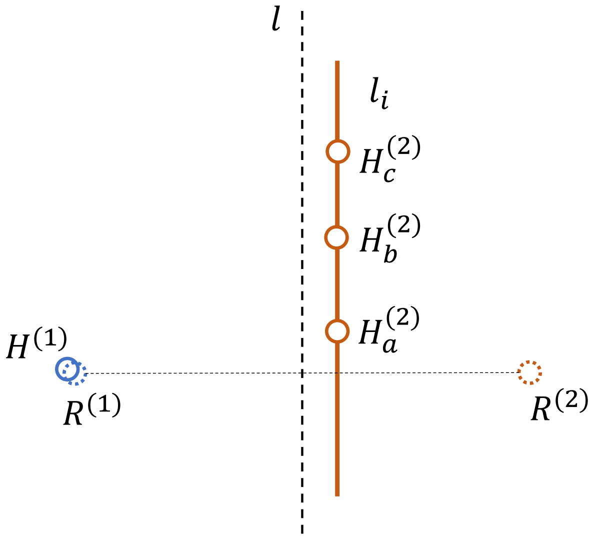


Figure 4: Experimental setup: how to find multiple options of $H^{(2)}$ for comparison? See Section 4.3.1.

Next, we will introduce how to find the equidistant hyperplane l and parallel hyperplanes $\{l_i\}$.

4.3.2 Find the equidistant hyperplane. The equidistant hyperplane is represented by

$$l = \{\vec{p} = \langle p_1, \dots, p_n \rangle \in \mathcal{R}^n | a_1p_1 + a_2p_2 + \cdots + a_np_n + c = 0\},$$

where c can be any nonzero constant. We are able to find the approximated values of the coefficients $a_1, \dots, a_n$ by randomly sampling n points on or extremely close to l_1 and solving a system of linear equations.

4.3.3 Find parallel hyperplanes. Similarly, the points on hyperplane parallel to the equidistant hyperplane are likely to satisfy our requirement. Formally, they are represented by

$$l_i = \{\vec{p} \in \mathbb{R}^n | \text{dist}(\vec{p}, l) = d\}$$

where the distance between $\vec{p} = \langle p_1, \dots, p_n \rangle$ and l can be calculated by

$$\text{dist}(\vec{p}, l) = \frac{a_1p_1 + a_2p_2 + \cdots + a_np_n + c}{\sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}}.$$

We may also say the distance between l and l_i is d.

4.3.4 Generation method. We use BART models with VAE ($l_z = 8$). We use all multi-referenced data for training and take out a subset as validation set to generate responses and evaluate at each training epoch. Specifically, we use ComVE, where each data instance has three references. We train the models with $R^{(1)}$ and $R^{(2)}$ for 50 epochs. We then add $R^{(3)}$ back in and finetune the models for another 5 epochs. We chose to do this so that our model can generate hypotheses close enough to each reference that is provided during training while being capable of generating diverse responses on the hyperplanes (with VAE framework).

For each input context, we generate 450 hypotheses (may include duplicates) and group them together based on their distance. Points with distance $d \pm \epsilon$ away from l are considered to be in the same group ($\epsilon = 0.025$).

4.3.5 Data Selection. We wish that hypotheses in a set are not too similar; otherwise, it is too hard for human annotators to rate and rank them. Besides, we hope that those hypotheses are not all incorrect generation, in which case accuracy and diversity will become secondary. Therefore, we apply filtering using BLEU and Self-BLEU of each set. More specifically, we select the sets of hypotheses with BLEU above the average BLEU of all sets and Self-BLEU below the average for human study.

4.4 Human Study Design

4.4.1 Criteria for humans to give ratings. In the experiment, we will collect ratings on the following criteria from human judges:

- Accuracy: Does the utterance accurately fit in the context?
- Diversity: How different or diverse is this generation given the previous generations?
- Overall: How do you judge the overall quality of the generation?

For hypotheses in each group, we ask human annotators to rate them under the RankME framework [17]. Specifically, we will put all hypotheses in a set in parallel and ask human annotators to rate each hypothesis from the three criteria separately. References will be set as a standard score 100 on each criterion. For example, hypotheses worse than references with respect to one criterion will be rated at a score below 100.

4.5 Initial Results

In Table 2, we include a subset of an example in ComVE dataset. Within the same hypothesis set, some achieve high diversity but give the wrong explanation, which is not desired. Other responses trade diversity to give a more accurate explanation.

From initial human studies, we observe that humans tend to value accuracy more than diversity, when the hyperplane is close to a reference, such as hypothesis set 3. However, for hypotheses sets that are close to the equidistant hyperplane (sets that are farther from the references), such as set 1, diversity is more preferred and the best response is the one that gives a relatively high diversity.

Table 2: An example in the ComVE dataset. The hypotheses in each set are ranked from the highest diversity to the lowest. Then naturally, as we demonstrated in Figure 4, the accuracy decreases from the top to the bottom.

Input Context
We should not help the weak.
References
Everyone has a hard time so we should help the weak. The weak need help. The biggest reward comes in helping the weak.
Hypothesis Set 1: $d = 0.1$
• The weak need to be treated with love and respect. *(best)* • The weak need to be treated by the best doctors and nurses. • The weak need to be protected. We should protect them. • We should help the weak they need help.
Hypothesis Set 2: $d = 0.26$
• The weak need help to survive. We should help them.. • The weak need to be strong. We should help them.
Hypothesis Set 3: $d = 0.48$
• The weak need help. We should not help the weak.*(wrong)* • The weak need help and we should help the strong. *(wrong)* • The weak need to be treated by the best.*(best)*

5 FUTURE DIRECTIONS

5.1 Designing Evaluation Metrics of Overall Semantic Quality

Although automatic accuracy metrics have shown positive correlations with human judgments, there is a noticeable gap between the existing diversity metrics and human evaluation in terms of semantic diversity [23]. Therefore, neural-based diversity metrics are of high demand and should take into account generation quality.

There is unlikely that a generalized metric can fit human judgments for all tasks. Indeed, some tasks such as story generation accept a higher level of diversity than tasks such as machine translation. Therefore, the new metrics should consider the types of text generation tasks.

5.2 Benchmarking with Human Judgement

Our proposed approach will perform human study to evaluate how well new metrics align with human judgment when there is a tradeoff between diversity and accuracy. However, human study is expensive and takes long time. To save the community's separated efforts, we are considering to develop a benchmark dataset with human judgement scores for evaluating new metrics about both accuracy and diversity.

6 CONCLUSIONS

In this paper, we conducted experiments to validate the trade-off between semantic accuracy and diversity on multi-reference text generation tasks beyond n-grams. The observation is consistent on different generation tasks and decoding methods. Furthermore, we proposed a novel human study framework and data collection methods to learn how humans make decisions on the tradeoff. Future work can utilize this framework to develop or evaluate new metrics as well as new benchmarks, eventually seeking to optimize existing models to generate both accuracy and diverse outputs.

ACKNOWLEDGMENTS

This work was supported in part by NSF IIS-1849816, IIS-2119531, IIS-2137396, IIS-2142827, CCF-1901059, and ONR N00014-22-1-2507.

REFERENCES

[1] Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly Measuring Diversity and Quality in Text Generation Models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Association for Computational Linguistics.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation.*

[3] Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. Language GANs Falling Short. arXiv:1811.02549 [cs.CL]

[4] Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture Content Selection for Diverse Sequence Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[5] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A. Young, and Brian Belgodere. 2022. Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge. *J. Artif. Int. Res.* 73 (2022).

[6] Yao Dou, Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2021. MultiTalk: A Highly-Branching Dialog Testbed for Diverse Conversations. In *AAAI Conference on Artificial Intelligence (AAAI).*

[7] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *AAAI Conference on Artificial Intelligence (AAAI).*

[8] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference for Learning Representation (ICLR).*

[9] Daphne Ippolito, Reno Kriz, Joao Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of Diverse Decoding Methods from Conditional Language Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL).*

[10] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference for Learning Representation (ICLR).*

[11] Marie-Anne Lachaux, Armand Joulin, and Guillaume Lample. 2020. Target Conditioning for One-to-Many Generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP-Findings).*

[12] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL).*

[13] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT).*

[14] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[15] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out.*

[16] Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate Your Counterfactuals: Towards Controlled Counterfactual Generation for Text. In *AAAI.*

[17] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable Human Ratings for Natural Language Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).*

[18] 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 (ACL).*

[19] Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 671–688.

[20] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL).*

[21] Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International Conference on Machine Learning (ICML).*

[22] Xinyao Shen, Jiangjie Chen, Jiaze Chen, Chun Zeng, and Yanghua Xiao. 2022. Diversified Query Generation Guided by Knowledge Graph. In *ACM Conference on Web Search and Data Mining (WSDM).*

[23] Guy Tevet and Jonathan Berant. 2021. Evaluating the Evaluation of Diversity in Natural Language Generation. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL).*

[24] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *AAAI Conference on Artificial Intelligence (AAAI).*

[25] Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 Task 4: Commonsense Validation and Explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation (SemEval-14).*

[26] Tianming Wang and Xiaojun Wan. 2019. T-CVAE: Transformer-Based Conditioned Variational Autoencoder for Story Completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19.* International Joint Conferences on Artificial Intelligence Organization.

[27] Wenhao Yu, Chenguang Zhu, Tong Zhao, Zhichun Guo, and Meng Jiang. 2021. Sentence-Permuted Paragraph Generation. In *Conference on empirical methods in natural language processing (EMNLP).*

[28] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc.

[29] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference for Learning Representation (ICLR).*

[30] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems (NeurIPS).*

[31] 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 *ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR).*