Variance-Aware Machine Translation Test Sets

Runzhe Zhan∗ Xuebo Liu∗ Derek F. Wong† Lidia S. Chao
NLP2CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.{runzhe,xuebo}@gmail.com, {derekfw,lidiasc}@um.edu.mo

Abstract

We release 70 lightweight and discriminative test sets for machine translation (MT) evaluation called variance-aware test sets (VAT), covering 35 translation directions from WMT16 to WMT20 competitions. VAT is automatically created by a novel variance-aware filtering method that filters the indiscriminative test instances of the current MT benchmark without any human labor. Experimental results show that VAT outperforms the original WMT benchmark in terms of the correlation with human judgement across mainstream language pairs and test sets. Further analysis on the properties of VAT reveals the challenging linguistic features (e.g., translation of low-frequency words and proper nouns) for the competitive MT systems, providing guidance for constructing future MT test sets. The test sets and the code for preparing variance-aware MT test sets are freely available at https://github.com/NLP2CT/Variance-Aware-MT-Test-Sets.

1 Introduction

The automated machine translation (MT) evaluation relies on the metrics and the test sets. Based on the usage of test sets, the metrics to quantify the performance of MT systems can be divided into two categories: reference-based metrics (Papineni et al., 2002; Popovic, 2015; Lo, 2019; Zhang et al., 2020) and reference-free metrics (Popovic, 2012; Yankovskaya et al., 2019). Among them, the reference-based metrics which measure the overlap between the reference and model’s hypothesis, are widely used both in research and practice. Even the state-of-the-art metrics that exploits the pre-trained model (Zhang et al., 2020; Sellam et al., 2020; Rei et al., 2020) are able to evaluate the finer-grained semantic overlap, it still cannot achieve human-level judgements (Ma et al., 2019; Mathur et al., 2020). Although the metric itself can be further elaborated, the reference in the test set, which is another key ingredient in the MT evaluation, receives less attention from the community.

The references are not innocent in confusing the automatic metrics. The existing research has proven that the collected references tend to exhibit the monotonic translation style (Popovic, 2019; Freitag et al., 2020a, 2020b) instead of natural text, which lack diversity in the evaluation. On the other hand, the competitive MT systems typically share a homogeneous architecture and training data, causing that the performance of the MT systems is too close to distinguish, thus the score divergence given by the automatic metrics becomes small. To alleviate this issue, previous work focused on increasing the reference diversity with the mean of paraphrasing, including human paraphrasing (Freitag et al., 2020a, 2020b) and automatic paraphrasing (Kauchak and Barzilay, 2006; Guo and Hu, 2019; Bawden et al., 2020), but both of them are expensive in terms of human labor and computational cost. Considering the fact that not all the references are monotonic, it is still unclear how to select those discriminative references instead of diversifying them for the MT evaluation.

This paper aims to tackle this issue without any human labor. Our motivation comes from a common fact in the real world. For the entrance examination, the simplest and most difficult questions

∗ Equal contribution
† Corresponding author

cannot tell the difference between the examinees because they may be all correctly or incorrectly answered. Accordingly, those questions that receive diverse answers play a vital role in distinguishing the examinees’ ability by comparing them with the ground truth. Based on this fact, a similar phenomenon may also happen in the test set of machine translation evaluation.

In this paper, we use the variance of translation scores evaluated by the metric as a criterion to create a variance-aware test set whose references are more discriminative in evaluating the MT systems. The selected references are characterized by their diverse evaluation scores, indicating that the MT systems are not consistent in translating the same source, thus this translation case is a valuable indicator in distinguishing the capability of the MT systems. Experimental results show that evaluating with the created variance-aware test set can improve the correlation with human judgements. Further analysis of the properties of the variance-aware test set also confirms its effectiveness.

Our main contributions are as follows:

- We release 70 variance-aware MT test sets, covering 35 translation directions from the WMT16 to WMT20 competitions. The test set filters 60% of test instances from the original WMT version, which is lightweight for research costing high computational resources (e.g., reinforcement learning and neural architecture search).
- We propose a simple and effective method to automatically identify discriminative test instances from MT test sets. We demonstrate that using the discriminative test instances can achieve a better correlation with human judgements than using the original test set.
- We give an in-depth analysis of the properties of discriminative and indiscriminative test instances. We find that the translations of low-frequency words and proper nouns are highly discriminative, providing clues for building challenging MT test sets.

2 Background

2.1 MT Evaluation and Meta-Evaluation

Machine translation evaluation is a crucial topic in MT development due to the demand for comparing the performance of several candidate MT systems. Traditionally, the human assessment is used to evaluate the MT systems but is expensive in terms of its costs. Moreover, the assessment quality of crowdsourced evaluation work is unpredictable, and there is a big gap between non-expert and professional translators (Toral et al., 2018; Läubli et al., 2020; Mathur et al., 2020). Therefore, the research on automatic evaluation metric receives a lot of attention due to its advantages such as low-cost and controllable process, which is now widely used in model selection and optimization (Shen et al., 2016; Wieting et al., 2019).

The reference-based metrics which rely on the reference translation are the most popular automatic evaluation metrics, and they differ in the measured overlap types. Specifically, BLEU (Papineni et al., 2002) and its variants (Doddington, 2002; Popović, 2015) evaluate the overlap by matching the n-gram, and other metrics like TER (Snover et al., 2006) quantify the overlap by the edit distance. However, these metrics are conducted in a hard matching paradigm and do not consider semantic. METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014) alleviates this issue by introducing synonymy and other linguistic features in word matching but is limited in the availability of language resources. Recent embedding-based metrics break the limitation of hard matching, making evaluate the semantic overlap possible. By enhancing the semantic representation by the pre-trained model (Devlin et al., 2019; Lample and Conneau, 2019), BERTScore (Zhang et al., 2020) correlates better with human judgements than previous metrics. At the same time, the end-to-end paradigm utilizing the pre-trained representation also is applied to the MT evaluation and achieves remarkable performance such as COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020).

To verify the effectiveness of automatic evaluation results, the process that measuring the correlation between scores given by an automatic metric and human ratings is called meta-evaluation. The meta-evaluation mainly uses the correlation coefficients like Pearson’s r to show how well the automatic metric performs like the human evaluator (Callison-Burch et al., 2006, 2008). The validation process of our method covers the mainstream metrics and uses the ordinary meta-evaluation methods to testify the correlation improvement.
2.2 Shortcoming of Current Test Sets

The less discriminative instances in the public benchmark are the bottleneck of automatic evaluation. The test sets released by the organizers of the WMT competition are the well-recognized benchmark for MT evaluation. However, some researchers argue that some references in these test sets may mislead the reference-based metrics, making the evaluation results of automatic metrics different from human judgement. One major question is that the existing references tend to be monotonic (Popovic, 2019; Freitag et al., 2020b). This translation style is easy to achieve by the MT systems and less discriminative for the evaluation. In addition, based on the phenomenon observed by Zhan et al. (2021) in the evaluation of WMT19 English→German task, most tokens can be correctly translated by all the participation systems, especially for those competitive ones, indicating that the test sets are partially valuable in distinguishing the MT systems.

Existing solutions are creating a more diverse test set to improve discernment. Kauchak and Barzilay (2006) firstly explored the automatic paraphrasing techniques in improving the accuracy of automatic metrics and validate the effectiveness on small-scale human assessment data. Bawden et al. (2020) further investigated the use of automatic paraphrasing in the automatic evaluation, showing the limited gains in correlation with human judgements on the WMT19 benchmark. Promisingly, human paraphrased references have been proved that it can significantly improve the correlation performance of BLEU on some language pairs (Freitag et al., 2020b). Overall, these methods to augment the references are restricted in the construction costs and consistent improvement.

Instead of diversifying the references, our work pays attention to selecting the discriminative part from the existing test set for better distinguishing among the strong MT systems.

3 Variance-Aware Test Set

In the educational entrance examination, the questions used to distinguish subjects may not be the most difficult. It is common sense that test-takers would not perform very differently when answering the simple questions, but the extremely difficult questions that no one can answer also cannot tell the differences of test-takers in terms of the ability.

This fact can also be mapped into the machine translation evaluation. As illustrated in Figure 1, evaluating the MT systems’ performance by using the first two references causes the subtle difference in evaluation results due to the polarized difficulty, thus it is hard to distinguish the systems’ capability in this circumstance. By contrast, the gap of evaluation results in the last case is huge enough to tell the differences in translation ability.

The cases clearly reveal that a discriminative test instance must make the evaluation exhibits large diversity so that it can become a decisive clue to compare the MT systems. Since the metrics evaluate the performance of MT systems, the discernment of test instances can be quantified by the variance of scores given by a metric, reflecting the degree of evaluation diversity. The higher variance indicates that using this test instance in the evaluation is easier to distinguish the systems. Therefore, our goal

![Figure 1: An illustration of proposed variance-aware filtering method.](image-url)
is to create a discriminative test set for better evaluating the MT systems by selecting the instances whose variance of evaluation results is high. This process is named variance-aware filtering.

3.2 Variance-Aware Filtering

To measure how MT systems perform differently towards a test instance, the performance of candidate systems is firstly quantified by the automatic metrics, then the standard deviation is simply used as a statistical indicator to model the diversity of evaluated performance. The standard deviation takes the square root of the variance, we use it because the scale of this measurement is the same as the original data. Given $N$ references $t = \{t_1, t_2, \ldots, t_N\}$ and a set of corresponding hypotheses $h = \{h_1, h_2, \ldots, h_N\}$ generated by $k$ systems in which $h_i = \{h_i^{(1)}, h_i^{(2)}, \ldots, h_i^{(k)}\}$, the performance diversity of hypothesis $h_i$ is estimated by the standard deviation $\sigma_i$ of scores, which is formulated as:

$$
\sigma_i = \sqrt{\frac{1}{k} \sum_{j=1}^{k} (M(h_i^{(j)}, t_i) - \mu_i)^2}, \quad 1 \leq i \leq N \tag{1}
$$

where $M(\cdot, \cdot)$ is the metric used to score the translation performance and $\mu_i$ is the average value of all the systems’ scores, which can be calculated as follows:

$$
\mu_i = \frac{1}{k} \sum_{j=1}^{k} M(h_i^{(j)}, t_i), \quad 1 \leq i \leq N \tag{2}
$$

For all the standard deviation $\{\sigma_1, \sigma_2, \ldots, \sigma_N\}$, higher $\sigma_i$ indicates that more differently the systems behave with respect to the reference $t_i$. Therefore, $\lambda$ percent of test instances whose corresponding references have lower $\sigma$ values will be filtered for creating a new discriminative test set, where $\lambda$ is a hyperparameter searched on the empirical experiments.

4 Experiments and In-Depth Analysis

4.1 Experimental Setup

Data  Five WMT test sets (Bojar et al., 2016, 2017; Ma et al., 2018, 2019; Mathur et al., 2020) ranging from WMT16 to WMT20 were used to conduct the experiments since it is the well-recognized benchmarks in the MT community, the included translation directions are as shown in Table 1. On the other hand, we choose the test sets starting from WMT16 because the neural machine translation (Bahdanau et al., 2015; Sennrich et al., 2016; Vaswani et al., 2017) paradigm has been largely improved the capability of MT systems and the systems submitted to the WMT competitions gradually become more competitive since 2016. For each language pair, the test set for meta-evaluation consists of references, system hypotheses, and corresponding human judgements.

Table 1: Detailed information of the test sets involved in the experiments where Num denotes the number of translation directions.

<table>
<thead>
<tr>
<th>WMT16 (Num=7)</th>
<th>WMT17 (Num=14)</th>
<th>WMT18 (Num=14)</th>
<th>WMT19 (Num=18)</th>
<th>WMT20 (Num=17)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>X-English</strong></td>
<td>cs, de, fi, ro, ru, tr, zh</td>
<td>cs, de, fi, lv, ru, tr, zh</td>
<td>de, fi, gu, kk, lt, ru, zh</td>
<td>cs, de, ru, ja, km, pl, ps, ru, ta, zh</td>
</tr>
<tr>
<td><strong>English-X</strong></td>
<td>ru, tr, zh</td>
<td>ru, tr, zh</td>
<td>cs, de, ja, pl, ru, ta, zh</td>
<td></td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td>ru, tr, zh</td>
<td>ru, tr, zh</td>
<td>cs, de, ja, pl, ru, ta, zh</td>
<td></td>
</tr>
</tbody>
</table>

Metrics and Meta-Evaluation  Without losing the generality, we validate our research hypotheses on the five representative metrics and use their publicly open-sourced implementations for easily reproducing the results:

- **BLEU** (Papineni et al., 2002) is an $n$-gram based metric that uses the precision rate to evaluate the coverage of reference $n$-gram in the model hypothesis. We use the sentence-level BLEU in the filtering procedure and evaluate the corpus-level system performance.
• **COMET** (Rei et al., 2020) is an end-to-end metric that builds on the top of the pre-trained XLM model including reference-based models and reference-free models. We use the recommended reference-based estimator model in the experiments.

• **BLEURT** (Sellam et al., 2020) is an end-to-end metric that fine-tunes the BERT model with several regression and classification tasks to make the model better adapt to the MT evaluation scenario. We use its released checkpoint and default settings in the experiments.

• **BERTScore** (Zhang et al., 2020) is an embedding-based metric that relies on the pre-trained BERT model to encode the reference and hypothesis, measuring the similarity of representation with precision (BERTS-P), recall (BERTS-R), and $F$-measure (BERTS-F). For the evaluation of different language pairs, we use the default BERT-family models as the same as the BERTScore implementation.

To examine the effectiveness of automatic evaluation metrics, we use the system-level Pearson’s $r$, Kendall’s $\tau$ and Spearman’s $\rho$ correlation coefficients as the criteria to measure how the automated evaluation results correlate with human judgements, which are also widely used in the competitions (Macháček and Bojar, 2013; Mathur et al., 2020) and related research (Freitag et al., 2020b).

### 4.2 Ablation Study

There are two main factors that may affect our proposed filtering approach: filtering percentage $\lambda$ and filtering metric $M$. Hence, a series of empirical experiments was conducted on the WMT20 benchmark to explore the best applicable settings to build the most discriminative test sets, and the finalized setting subsequently would be used to validate its generality in other WMT benchmarks.

**Choice of Filtering Percentage** $\lambda$ affects the amount of data to be reserved and also is an indicator that reflects the discernment of the current data sets with respect to the evaluation metrics. As shown in Figure 2, only using the partial test set can improve the evaluation correlation of automatic metrics, but the most effective percentage setting depends on the type of evaluation metrics. Compared to BLEU metric, the metrics driving by the pre-trained models achieve the local optimal correlation using less proportion of the test set, i.e., $\lambda \geq 50$. The underlying reason may lie in the evaluation granularity in terms of the semantics, the metric which is better in parsing the semantic needs fewer data to distinguish the MT systems because the impact brought by those discriminative samples in the comparison is larger. For achieving the better correlation results and fitting most of the metrics, we filter 60% instances out of the original test sets in the subsequent experiments.

**Choice of Filtering Metric** $M$ matters because the discernment of test instances can not be estimated without the accurate evaluation of the MT systems’ performance. Figure 3 presents how the scores given by the different metrics affect the correlation of filtered test sets. Filtering the test set based on the scores given by the BERTS-R metric outperforms the test sets created by other metrics. It is reasonable that the BERTS-R metric consistently achieves the best correlation when using it as the evaluation metric (also as shown in Figure 2), thus is better to quantify the differences between
Figure 3: Comparison of averaged correlation results measured on all the WMT20 translation directions under different filtering metrics. Filtering the test sets by BERTS-R scores consistently achieve stable correlation results across different evaluation metrics. Using COMET scores has comparable correlation performance with BERTR-S except Pearson correlation results.

Table 2: Comparison of averaged correlation results using original and variance-aware test sets (VAT) where Num denotes the number of language pairs. Evaluating MT systems with variance-aware test sets (+VAT) better correlates with human judgements across different evaluation metrics.

<table>
<thead>
<tr>
<th>Metric</th>
<th>WMT16 (Num=7)</th>
<th>WMT17 (Num=14)</th>
<th>WMT18 (Num=14)</th>
<th>WMT19 (Num=18)</th>
<th>WMT20 (Num=17)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( r )</td>
<td>( \tau )</td>
<td>( \rho )</td>
<td>( r )</td>
<td>( \tau )</td>
</tr>
<tr>
<td>BLEU</td>
<td>.826</td>
<td>.645</td>
<td>.778</td>
<td>.910</td>
<td>.737</td>
</tr>
<tr>
<td>+VAT</td>
<td>.880</td>
<td>.723</td>
<td>.837</td>
<td>.928</td>
<td>.754</td>
</tr>
<tr>
<td>COMET</td>
<td>.988</td>
<td>.886</td>
<td>.958</td>
<td>.982</td>
<td>.882</td>
</tr>
<tr>
<td>+VAT</td>
<td>.988</td>
<td>.881</td>
<td>.955</td>
<td>.985</td>
<td>.885</td>
</tr>
<tr>
<td>BLEURT</td>
<td>.982</td>
<td>.856</td>
<td>.942</td>
<td>.939</td>
<td>.789</td>
</tr>
<tr>
<td>+VAT</td>
<td>.984</td>
<td>.859</td>
<td>.944</td>
<td>.951</td>
<td>.807</td>
</tr>
<tr>
<td>BERTR-S</td>
<td>.976</td>
<td>.848</td>
<td>.932</td>
<td>.947</td>
<td>.866</td>
</tr>
<tr>
<td>+VAT</td>
<td>.976</td>
<td>.840</td>
<td>.938</td>
<td>.960</td>
<td>.820</td>
</tr>
<tr>
<td>BERTR-F</td>
<td>.941</td>
<td>.831</td>
<td>.931</td>
<td>.974</td>
<td>.825</td>
</tr>
<tr>
<td>+VAT</td>
<td>.953</td>
<td>.854</td>
<td>.943</td>
<td>.972</td>
<td>.826</td>
</tr>
<tr>
<td>BERTR-P</td>
<td>.975</td>
<td>.881</td>
<td>.956</td>
<td>.970</td>
<td>.833</td>
</tr>
<tr>
<td>+VAT</td>
<td>.979</td>
<td>.900</td>
<td>.964</td>
<td>.974</td>
<td>.842</td>
</tr>
</tbody>
</table>

4.3 Main Results

By applying the filtering settings explored in the previous sections to other WMT benchmarks, Tables 2 and 3 presents the comparison of correlation results between using the filtered and original test sets. The improved correlation performance across most metrics and benchmarks consistently confirms that the effectiveness of evaluating with variance-aware test set (VAT), especially for the metrics powered by the pre-trained models. As for the n-gram-based metrics, it may over-penalize the overlap that shares the same semantic due to the hard-matching paradigm, making some VAT instances inactive in evaluating the diverse hypotheses. In contrast to the hard-matching paradigm, the metrics utilizing the pre-trained models are able to fairly judge synonymous expressions, thus the created VAT are substantially useful to distinguish the MT systems.

4.4 Analysis of Variance-Aware Test Sets

To investigate how the correlation improvement benefits from VAT, we characterize VAT build on the WMT20 benchmark from the perspective of the linguistic and data properties in this section.
## Table 3: Comparison of Pearson correlations using original and variance-aware test sets (VAT) on some mainstream language pairs. T. denotes the WMT test set. Using variance-aware test sets (+VAT) consistently improves the evaluation results of the language pairs across different test sets.

<table>
<thead>
<tr>
<th>Metric</th>
<th>De-En</th>
<th>En-De</th>
<th>Zh-En</th>
<th>En-Zh</th>
<th>En-Cs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>.928</td>
<td>.969</td>
<td>.888</td>
<td>.819</td>
<td>.980</td>
</tr>
<tr>
<td>+VAT</td>
<td>.940</td>
<td>.975</td>
<td>.925</td>
<td>.845</td>
<td>.981</td>
</tr>
<tr>
<td>COMET</td>
<td>.963</td>
<td>.997</td>
<td>.904</td>
<td>.797</td>
<td>.987</td>
</tr>
<tr>
<td>+VAT</td>
<td>.979</td>
<td>.998</td>
<td>.944</td>
<td>.841</td>
<td>.987</td>
</tr>
<tr>
<td>BERTS-P</td>
<td>.948</td>
<td>.998</td>
<td>.947</td>
<td>.798</td>
<td>.988</td>
</tr>
<tr>
<td>+VAT</td>
<td>.964</td>
<td>.999</td>
<td>.952</td>
<td>.830</td>
<td>.989</td>
</tr>
<tr>
<td>BERTS-F</td>
<td>.938</td>
<td>.997</td>
<td>.946</td>
<td>.939</td>
<td>.990</td>
</tr>
<tr>
<td>+VAT</td>
<td>.981</td>
<td>.999</td>
<td>.952</td>
<td>.876</td>
<td>.989</td>
</tr>
</tbody>
</table>

### Figure 4: Absolute constitution changes of variance-aware test sets in terms of sentence length. VAT reserved more short sentences for most language pairs. The boundaries for splitting the group are determined by the equidistant percentile of sentence length.

**Sentence Length** generally associates with the translation difficulty [Koehn and Knowles, 2017], but the difficult sentence may be less relevant to the high discernment. As shown in Figure 4, longer sentences exhibit lower discernment and were filtered by our method. Translating longer sentences is extremely challenging for MT systems due to the long-distance dependency or complex entity relationships [Cho et al., 2014, Sennrich and Haddow, 2016, Eriguchi et al., 2019], leading to close translation performance of MT systems. On the contrary, short sentences are more discriminative because different systems tend to show greater differences in terms of syntactic and lexical choices. But some special translation directions show the opposite trends such as Chinese→English and German→English. Not only the number of systems that participated in these translation tasks is relatively huge, but also the systems trained on these high-resource language pairs are like to be more competitive. Since the competitive systems can be good at translating the short sentence, the clues to judge their capability could rely on the translation of medium or long sentences.

**Word Frequency** is a measurement that reflects the finer-grained difference of MT systems since they may vary in the lexical choice of rare words [Koehn and Knowles, 2017, Ding et al., 2021b]. As shown in Figure 5, the proportion of frequently occurred words in the training set is reduced by in the VAT, indicating that high-frequency words are less discriminative. The representations of high-frequency words learned on the training set tend to be stable, whereas the low-frequency words
are insufficiently learned. Particularly, some systems may enhance the translation performance of low-frequency words with the help of data augmentation (Fadaee et al., 2017; Ding et al., 2021a) or representation enhancement techniques (Nguyen and Chiang, 2018; Liu et al., 2019), resulting in the differences of lexical choice performed on the test set. Overall, the percentage change of word frequencies is not so large as the sentence length comparison because the filtering operation is conducted at the sentence level, thus only those sentences whose proportion of low-frequency words is high will be reserved.

Figure 5: Absolute constitution changes of variance-aware test sets in terms of word frequency. VAT filtered the sentences which contain more frequent words. The boundaries for categorizing the “Rare”, “Middle”, “Frequent” group are 20%, 60%, 100% percentile of word frequency, respectively.

Figure 6: Absolute constitution changes of variance-aware test sets in terms of English part-of-speech tagged by the NLTK toolkit. VAT has more proper nouns than the original test sets.

Part-of-Speech better depicts the lexical features of VAT considering the syntactic role a word plays. It can be seen from Figure 6 that VAT reserved more sentences containing the proper nouns (NNP). This phenomenon echoes our previous comparative exploration of word frequency since there is a large overlap between NNPs and low-frequency words like technical terms of a specific domain,
but the translation performance of NNP s is not as intractable as long-sentence translation. Due to the
fact that the bottleneck of the long-sentence translation may be related to the model architecture (Cho
et al., 2014), most MT systems that share the homogeneous architecture (Vaswani et al., 2017) still
are problematic in translating these challenging sentences. Similar to the problem of low-frequency
words, the poor translation accuracy of NNP s can be alleviated by introducing external knowledge
(Chatterjee et al., 2017) or domain adaptation techniques (Hu et al., 2019) concerning data-efficient
learning, thus evaluating the translation of NNP s are also valuable in distinguishing MT systems.

**Human Paraphrasing** is effective to improve the correlation of automatic metrics by divers-
sifying the reference but is expensive and not sustainable. However, there are some references
whose style is close to the human natural language that does not require many efforts in para-
phrasing. If our proposed method is able to select these references, the filtering procedure
ought to be similar to diversify the references without human labor. To verify this claim, we
calculate the edit distance (Levenshtein, 1966) between different test sets and human paraphrased
test sets provided by Freitag et al. (2020b). Table 4 clearly shows that our method does reserve the
references that are closer to the human paraphrased references compared to the original test set and
filtered data, and the deeper reason of correlation improvement brought by VAT may be similar to the
effect of human paraphrasing.

**Translation Difficulty** is not the same as the evaluation discernment as we assumed before. The test instances
with extreme translation difficulty are not applicable in the evaluation aiming to distinguish the MT systems, so
the filtered samples possibly are not the simplest or the most difficult instances. Starting from this intuition, we
investigate averaged score distribution of the test instances on the WMT20 English→German translation task since
it involves competitive systems that are challenging to the automated evaluation (Freitag et al., 2020b). Obviously,
Figure 7 reveals that the reserved instances have moderate but not extreme difficulty. The phenomenon that slightly
higher difficulty samples are reserved also conform to the previous observation in terms of the sentence length that
the longer sentences are more vital in distinguishing the MT systems due to their strong capability, also echoes
the previous research stating that this translation direction easily confused the automatic evaluation metric (Freitag et al., 2020b).

To conclude, the test item reserved by the proposed filtering method is discriminative in terms of
the linguistic and data properties, and the correlation improvement of the variance-aware test set is
reasonable. Moreover, the variance-aware filtering method has the potential in saving the human
labor for diversifying the test sets.

### 5 Conclusions and Future Work

This paper introduces a method to select those discriminative test instances from the machine transla-
tion benchmark and create a series of variance-aware test sets in an automatic way. Experimental
results show that using the created test sets can improve the correlation performance of automatic
evaluation results across the representative test sets and languages, confirming the effectiveness and
generality of the proposed method. Further analysis of the features of test instances supports the
rationality of variance-aware test sets and ensures its scalability for other possible usages.

Future work will investigate the use of the variance-aware test sets in other MT research questions like
architecture search, and extend the filtering method to other evaluation tasks like dialogue generation.

---

**Table 4:** Edit distances between different parts of English→German test set with corresponding human paraphrased data. The sentences reserved in the VAT is closer to the human paraphrased data.

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Filtered</th>
<th>Reserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT20</td>
<td>30.35</td>
<td>30.46</td>
<td>30.18</td>
</tr>
<tr>
<td>WMT19</td>
<td>19.82</td>
<td>20.25</td>
<td>19.17</td>
</tr>
<tr>
<td>WMT18</td>
<td>20.02</td>
<td>20.30</td>
<td>19.72</td>
</tr>
</tbody>
</table>

**Figure 7:** Comparison of distribution of BERTS-R scores between filtered and reserved sentences. Sentences with medium difficulty are more discrimina-
tive than those with extreme difficulty.

[Bojar et al., 2018] [Barrault et al., 2019].
References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] See Appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 4.4
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 4.2, 4.3 and 4.4

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the URL listed in the abstract.
(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.2.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] The results will not be affected by different random seeds.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A] The data released for the WMT news translation task can be freely used for research purposes. The details can be accessed in the official websites of the WMT competition, e.g., [http://www.statmt.org/wmt20/translation-task.html].

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] Similar to 4(d).

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We used the human judgements data released by the organizer of WMT competitions following the license of research usage.