Hacking Neural Evaluation Metrics with Single Hub Text

Hiroyuki Deguchi[†] **Katsuki Chousa**[†] **Yusuke Sakai**[‡]
[†]NTT, Inc. [‡]Nara Institute of Science and Technology
{hiroyuki.deguchi,katsuki.chousa}@ntt.com sakai.yusuke.sr9@is.naist.jp

Abstract

Strongly human-correlated evaluation metrics serve as an essential compass for the development and improvement of generation models and must be highly reliable and robust. Recent embedding-based neural text evaluation metrics, such as COMET for translation tasks, are widely used in both research and development fields. However, there is no guarantee that they yield reliable evaluation results due to the black-box nature of neural networks. To raise concerns about the reliability and safety of such metrics, we propose a method for finding a single adversarial text in the discrete space that is consistently evaluated as high-quality, regardless of the test cases, to identify the vulnerabilities in evaluation metrics. The single hub text found with our method achieved 79.1 COMET% and 67.8 COMET% in the WMT'24 English-to-Japanese (En–Ja) and English-to-German (En–De) translation tasks, respectively, outperforming translations generated individually for each source sentence by using M2M100, a general translation model. Furthermore, we also confirmed that the hub text found with our method generalizes across multiple language pairs such as Ja–En and De–En.

1 Introduction

Automatic evaluation metrics for measuring the quality of machine-generated content play a crucial role in improving generation models and must be highly reliable and robust. Recent embedding-based neural evaluation metrics, such as COMET (Rei et al., 2020, 2022) for translation tasks, have achieved high correlations with human assesments (Kocmi et al., 2021) and widely used in both research and development fields (Kocmi et al., 2024; Freitag et al., 2022, 2024; Rei et al., 2024; Fernandes et al., 2022) compared with previous lexical metrics, such as CHRF (Popović, 2015).

Nevertheless, there is no guarantee that these metrics yield reliable evaluation scores due to the black-box nature of neural networks. Zhang et al. (2025) created adversarial hubs in continuous space for images and audio by exploiting the hubness problem (Radovanović et al., 2010) in multi-modal retrieval tasks. Hub vectors are known to appear in high-dimensional continuous spaces and tend to be close to many examples. However, unlike images and audio, which can be searched via gradient descent, a hub text is more difficult to find because we need to search in a discrete space, i.e., NP-hard.

Hub text	podnikáníゴញ ウ»的聲音毎回強いメッセージを提供するかどうか模索まれたပေမယ့် ตอน ไปのこと歓通だった。
Source	But the new ones cover that area just fine.
Reference	でも、新しいやつはちょうどうまくその辺りをカバーしてる。
Score	COMET% = 91.7 CHRF% = 4.8
Source	It's a year of transition for me.
Reference	僕にとっては変化の年だ。
Score	COMET% = 89.6 CHRF% = 3.8

Table 1: Hub text of COMET and its evaluation scores in En–Ja translation. Even if other sources and their corresponding reference translations are given, hub text is evaluated with high COMET scores.

In this study, we first found that the hubness problem also occurs in automatic neural evaluation metrics operated in *discrete space* of text generation. Table 1 shows a hub text we discovered, which is consistently evaluated as high-quality on COMET (Rei et al., 2020, 2022), regardless of source and reference texts, in English-to-Japanese (En–Ja) translation. We propose a method for finding hub texts and attacking neural metrics to reveal their vulnerabilities. We first train a hub embedding in the hidden space by maximizing the evaluation score. We then decode it into its corresponding hub text using an inversion model (Morris et al., 2023). The obtained hub text appears to be a natural sentence, yet it already receives an inappropri-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 198–206

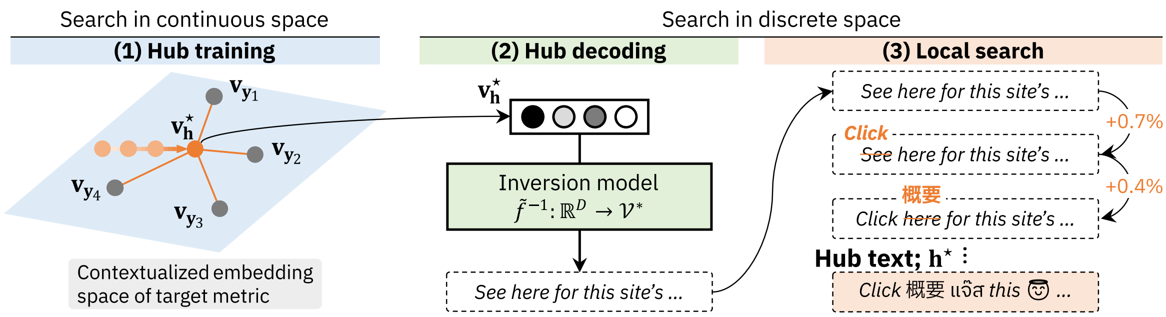

Figure 1: Overview of our proposed method for finding hub text

ately high score. To further exploit this issue, we search for an even more overestimated sentence. We then apply local search, a heuristic algorithm, to refine the hub text so that it maximizes the evaluation score. Table 1 shows cases that are assigned unreasonably high scores with evaluation metrics, thereby exposing serious concerns about their reliability and robustness. Indeed, some real-world situations rely solely on the COMET score, e.g., corpus filtering (Peter et al., 2023) and decision-making (see Section 5.1), resulting in the existence of hub text having practical impact.

We investigated a case study of COMET in the WMT'23 and WMT'24 En–Ja and En–De translation tasks (Kocmi et al., 2023, 2024). As a result, we observed that the single hub text found using our method achieved a higher COMET score than M2M100-generated translations (Fan et al., 2021), even though M2M100 translated for each source text. Furthermore, we also confirmed that the hub text generalizes across multiple language pairs such as Ja–En and De–En. Our findings highlight the necessity of using sanity checks or multiple metrics from the perspective of the hubness problem.

2 Background and Related Work

Embedding-based text evaluation metrics Automatic evaluation metrics assess the quality of machine-generated content. Recent embedding-based neural metrics, such as COMET (Rei et al., 2020, 2022) for machine translation tasks, have achieved high correlations with human assessments. They encode a hypothesis text into its embedding and calculate similarity with the input and/or reference. In this study, we primarily focus on translation tasks and use one of its embedding-based neural evaluation metrics, COMET, as a case study.

Let $\mathbf{x} \in \mathcal{V}^*$ and $\mathbf{y} \in \mathcal{V}^*$ be the input and reference output text, respectively, where $\mathcal{V}^*$ is the Kleene closure of vocabulary $\mathcal{V}$. COMET evaluates the hypothesis text $\mathbf{h} \in \mathcal{V}^*$, generated using a translation model, and calculates its quality score,

$$S(\mathbf{h}; \mathbf{x}, \mathbf{y}) \coloneqq s\left(\mathbf{v}_{\mathbf{x}}, \mathbf{v}_{\mathbf{h}}, \mathbf{v}_{\mathbf{y}}\right), \qquad (1)$$

where $\mathbf{v}_{\cdot} = f(\cdot)$, $f\colon \mathcal{V}^* \to \mathbb{R}^D$ is a sentence encoder, $s\colon \mathcal{V}^* \times \mathcal{V}^* \times \mathcal{V}^* \to \mathbb{R}$ is an output layer, and $D \in \mathbb{N}$ is the size of the embedding dimension.

Hubness Hubness (Radovanović et al., 2010) is a phenomenon in high-dimensional spaces, where hub embeddings frequently appear among the nearest neighbors of many examples, leading to irrelevant retrievals in information retrieval. While many studies aimed to mitigate this issue (Dinu et al., 2015; Wang et al., 2023; Chowdhury et al., 2024), we instead exploit it to uncover vulnerabilities in neural metrics. Zhang et al. (2025) created adversarial hubs for images and audio. These hubs exist in continuous space and can be easily optimized via gradient descent. In contrast, our goal is to find hub texts in the discrete text space, i.e., NP-hard.

3 Methodology

To identify the vulnerability of embedding-based neural metrics, we propose how to find the hub text $\mathbf{h}^{\star} \in \mathcal{V}^*$ that always receives a high score regardless of the input and reference text. Formally, our goal is to find the solution of the following objective:

$$\mathbf{h}^{\star} \coloneqq \operatorname{argmax}_{\mathbf{h} \in \mathcal{V}^*} \sum\nolimits_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}} S(\mathbf{h}; \mathbf{x}, \mathbf{y}), \qquad (2)$$

where $\mathcal{D} \subseteq \mathcal{V}^* \times \mathcal{V}^*$ is the test set that consists of pairs of the input and its reference text.

Our method finds the hub text through three steps: (1) hub training, (2) hub decoding, and (3) local search, as illustrated in Figure 1. We first train the optimal hub embedding in the contextualized embedding space of the target metric. The

Algorithm 1: Local search

Given : The score output layer s and the pairs of input and its reference output embeddings $(\mathbf{v_x}, \mathbf{v_y})$ calculated from $\mathcal{D}_{\text{tune}}$.
Input : The initial hub text $\mathbf{h}^{(0)} \in \mathcal{V}^*$.
Output : The refined hub text $\mathbf{h}^{\text{best}} \in \mathcal{V}^*$.

$S^{\text{best}} \leftarrow -\infty, t \leftarrow 0, \mathbf{h}^{\text{best}} \leftarrow \mathbf{h}^{(0)}$
repeat
    $t \leftarrow t + 1$
    for $i \leftarrow 1 \ldots |\mathbf{h}^{(t-1)}|$ **do**
        for each $v \in \mathcal{V}$ **do**
            // $\circ$ denotes concatenation.
            // $\mathbf{h}_{a:b}$ denotes the slice of $\mathbf{h}$ from a to b inclusive.
            $\mathbf{h}^{\text{cur}} \leftarrow \mathbf{h}^{\text{best}}_{1:i-1} \circ v \circ \mathbf{h}^{\text{best}}_{i+1:|\mathbf{h}^{\text{best}}|}$
            $\mathbf{v}_{\mathbf{h}^{\text{cur}}} \leftarrow f\left(\mathbf{h}^{\text{cur}}\right)$
            $S^{\text{cur}} \leftarrow 0$
            for $n \leftarrow 1 \ldots |\mathcal{D}_{\text{tune}}|$ **do**
                $S^{\text{cur}} \leftarrow S^{\text{cur}} + s(\mathbf{v}_{\mathbf{x}^{(n)}}, \mathbf{v}_{\mathbf{h}^{\text{cur}}}, \mathbf{v}_{\mathbf{y}^{(n)}})$
            if $S^{\text{cur}} > S^{\text{best}}$ **then**
                $S^{\text{best}} \leftarrow S^{\text{cur}}$
                $\mathbf{h}^{\text{best}} \leftarrow \mathbf{h}^{\text{cur}}$
    $\mathbf{h}^{(t)} \leftarrow \mathbf{h}^{\text{best}}$
until $\mathbf{h}^{(t)} = \mathbf{h}^{(t-1)}$
return $\mathbf{h}^{\text{best}}$

hub embedding is then decoded into its corresponding text. Finally, the hub text is refined through local search to maximize the evaluation score on the tuning data.

(1) Hub training This step finds the hub embedding $\mathbf{v}^{\star}_{\mathbf{h}} \in \mathbb{R}^D$ in the embedding space:

$$\mathbf{v}^{\star}_{\mathbf{h}} := \operatorname{argmax}_{\mathbf{v_h} \in \mathbb{R}^D} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{\text{tune}}} s(\mathbf{v_x}, \mathbf{v_h}, \mathbf{v_y}), \quad (3)$$

where $\mathcal{D}_{\text{tune}} \subseteq \mathcal{V}^* \times \mathcal{V}^*$ denotes the parallel data. To obtain the $\mathbf{v}^{\star}_{\mathbf{h}}$, we optimize $\mathbf{v_h}$ by treating it as learnable parameters. The other embedding vectors and layers, i.e., $\mathbf{v_x}$, $\mathbf{v_y}$, and s, are frozen.

(2) Hub decoding Next, we decode the hub embedding $\mathbf{v}^{\star}_{\mathbf{h}}$ into its corresponding text $\mathbf{h}^{\star}$. Since the previous step does not provide a concrete text, we need to find its corresponding hub text from the discrete space $\mathcal{V}^*$. Decoding a hub embedding is equivalent to applying the inverse function of the text encoder $f^{-1}\colon \mathbb{R}^D \to \mathcal{V}^*$, but non-linear encoders make it hard to exactly formulate the inverse function. Thus, we instead approximate it using an encoder-decoder inversion model $\tilde{f}^{-1}\colon \mathbb{R}^D \to \mathcal{V}^*$, which generates the original text $\mathbf{h}$ from its embedding $\mathbf{v_h}$ (Morris et al., 2023):

$$\hat{\theta} = \operatorname{argmax}_{\theta} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{mono}}} p_{\theta}(\mathbf{x} \mid f(\mathbf{x})), \quad (4)$$

where $\hat{\theta}$ denotes learned parameters, and $\mathcal{D}_{\text{mono}} \subseteq \mathcal{V}^*$ denotes a monolingual corpus.

Step	Hub text
(2) Hub decoding	『ザバラア』もそれまでどんな特徴を持つか悩んでいたのですが、それは予想に反していなかった。
(3) Local search	podnikáníゴញ ウ»的聲音每回強いメッセージを提供するかどうか模案まれたပေမယ့် ตอน ไป のこと歓通だった。

Table 2: Hub texts found for each search step. Note that step (1) does not identify concrete text, as it searches over continuous space in En–Ja.

During decoding, we generate multiple hypotheses of hub texts $\mathcal{H} := \{\mathbf{h}_i \sim p_{\hat{\theta}}(\mathbf{h}|\mathbf{v}^{\star}_{\mathbf{h}})\}_{i=1}^{|\mathcal{H}|} \subseteq \mathcal{V}^*$ and select the hypothesis that maximizes the score on the tuning data $\mathcal{D}_{\text{tune}}$, i.e., $\operatorname{argmax}_{\mathbf{h} \in \mathcal{H}} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{\text{tune}}} S(\mathbf{h}; \mathbf{x}, \mathbf{y})$. This strategy is similar to minimum Bayes' risk decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020), but we use true references of the tuning data $\mathcal{D}_{\text{tune}}$ instead of pseudo-references, which are typically generated from text generation models.

(3) Local search To further increase the score, we refine the hub text using local search, a heuristic search algorithm. Specifically, we iteratively replace tokens with those that maximize the score at each token position, as shown in Algorithm 1. We obtain the refined hub text $\mathbf{h}^{\text{best}}$ using the generated text in step (2) as the initial text $\mathbf{h}^{(0)}$.

4 Experiments

Setup We experimented on the En–Ja and En–De translation tasks using COMET[1]. We used WMT'23 (Kocmi et al., 2023) and WMT'24 (Kocmi et al., 2024) translation tasks for tuning and testing purposes, respectively. Through all experiments, for each language pair, we found a single hub text using WMT'23, and evaluated it on WMT'24. The details of these datasets are listed in Appendix B. We compared the evaluation scores with translations generated with M2M100, which has 418M parameters (Fan et al., 2021), with a beam size of 5. The hub embedding was initialized by averaging the embeddings of reference texts over the tuning data. We optimized the hub embedding using the AdamW optimizer (Loshchilov and Hutter, 2019) ($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10^{-8}$ and a weight decay with a coefficient 0.01) in 10,000 steps with a learning rate of 10^{-5}. For the in-

[1] Unbabel/wmt22-comet-da

Hypotheses	En–Ja WMT'23 (Dev) COMET	En–Ja WMT'23 (Dev) CHRF	En–Ja WMT'24 (Test) COMET	En–Ja WMT'24 (Test) CHRF	En–De WMT'23 (Dev) COMET	En–De WMT'23 (Dev) CHRF	En–De WMT'24 (Test) COMET	En–De WMT'24 (Test) CHRF
M2M100	78.6 ±12.8	24.6	71.4 ±15.4	21.4	66.0±15.8	51.8	66.0±16.7	46.6
Search in continuous space								
(1) Hub training	93.2 ±2.7	N/A	91.1 ±4.8	N/A	97.3 ±0.9	N/A	97.1 ±1.5	N/A
Search in discrete space								
(2) Hub decoding	65.9 ±7.2	5.6	61.3 ±7.8	4.2	46.9 ±4.8	13.5	46.6 ±5.0	16.3
(3) Local search	83.1 ±5.2	3.5	79.1 ±6.9	2.7	68.4±5.4	9.7	67.8±6.6	12.6

Table 3: COMET% with standard deviation and CHRF% scores of single hub text on COMET and translations generated using M2M100. We used WMT'23 as tuning set and WMT'24 as test set.

System	COMET%
ONLINE-B	88.2
ONLINE-W	87.5
ONLINE-Y	87.3
GPT4-5shot	87.0
SKIM	86.6
NAIST-NICT	86.2
ZengHuiMT	85.3
ONLINE-A	85.2
Lan-BridgeMT	84.5
ONLINE-M	13.3
Single hub text	83.1
ANVITA	82.7
KYB	80.8
AIRC	80.7
ONLINE-G	80.4
NLLB_Greedy	79.3
NLLB_MBR_BLEU	77.7

Table 4: Leaderboard of WMT'23 En–Ja translation task. These scores are cited from the shared task description paper (Kocmi et al., 2023). Unconstrained systems are indicated with gray background in tables, following Kocmi et al. (2023), and hub text found with our method is highlighted in green .

version model, we fine-tuned mT5-base (Xue et al., 2021). We used the random 1,000,000 Japanese sentences sampled from JParaCrawl v3 (Morishita et al., 2022) and parallel corpus from Common-Crawl provided on WMT'23 (Kocmi et al., 2023) for the training data of the En–Ja and En–De inversion model, respectively. In the step of hub decoding, we generated 1,024 hypotheses and selected one that maximizes the evaluation score on the tuning data.

Results Table 2 shows the hub texts found with our method for each search step in En–Ja. As shown in Table 2, step (2) found a hub text in natural language, and step (3) found a hub text that falls outside the bounds of natural language, yet maximizes the evaluation scores. Table 3 demonstrates the evaluation results. Step (1) achieved extremely high scores, 91.1% in En–Ja and 97.1% in En–De. This is because the hub embedding is optimized in a continuous space, but there may be no corresponding concrete text. Step (2) degraded from step (1), i.e., the scores were less than 70.0% in En–Ja, due to discretizing the hub embedding into the token sequence through the inversion model. Finally, step (3) improved the COMET scores by searching tokens that maximize the score over the vocabulary and achieved 79.1% in En–Ja and 67.8% in En–De. This step can generate texts beyond natural languages, as it systematically searches over the vocabulary, including low-frequency words. We also observed that the single hub text achieved higher COMET scores than M2M100's translations in both WMT'23 and WMT'24, despite having extremely low CHRF scores. In addition, the standard deviation (SD) of the COMET scores was lower than that of M2M100. These results indicate that hub texts consistently receive unreasonably high scores. Thus, we reveal that the existence of hub texts exposes critical vulnerabilities in COMET.

5 Discussion

5.1 Impact of existence of hub texts in real-world scenario

We show the leaderboard of the WMT'23 En–Ja translation task in Table 4. Note that this leaderboard is cited from the paper of the shared task (Kocmi et al., 2023). In addition, in the WMT'24 translation task, COMET is used to determine whether human evaluation is applied. Thus, the hub text can affect the evaluation of other translation systems on the leaderboard. Specifically, in WMT, only systems with high automatic evaluation scores are selected for human evaluation. In the case of Table 4, since the hub text achieved higher scores in COMET than ANVITA, KYB, and AIRC, it may hinder fair evaluation. Moreover, such auto-

	Ja–En		En–De		De–En	
Hypotheses	COMET	CHRF	COMET	CHRF	COMET	CHRF
M2M100	69.1	34.6	66.0	47.1	75.6	51.8
Hub text	63.4	1.3	62.1	0.4	60.7	0.6

Table 5: Evaluation results of hub text found using WMT'23 En–Ja in non-target language pairs, WMT'23 Ja–En, En–De, and De–En

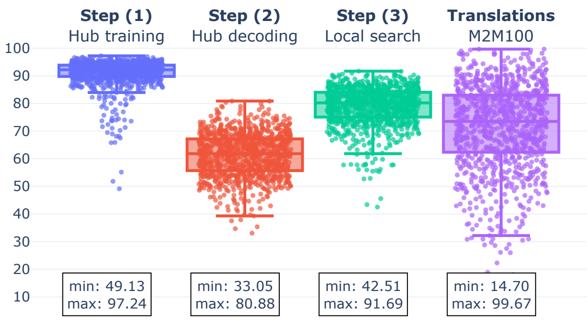

Figure 2: Scatter and box plots of COMET% scores for each test case in WMT'24 En–Ja.

matic filtering based on neural evaluation metrics is also applied in other contexts, e.g., corpus filtering. In such large-scale batch processing, COMET is often used as a reliable metric without human verification for each instance, and the presence of hub texts is already impactful in real-world scenarios.

5.2 Evaluation in non-target languages

To clarify whether the hub text is language-agnostic in multilingual neural evaluation metrics, i.e., whether the vulnerability is shared between other languages, we evaluated the hub text found using En–Ja in other language pairs. Table 5 shows the evaluation results of the hub text generated on the WMT'23 En–Ja in the WMT'23 Ja–En, En–De, and De–En translation tasks. Unlike in En–Ja, the COMET score of the hub text was lower than that of the translations generated using M2M100. However, the score still exceeded 60% and was not as low as 0.4 CHRF% observed in En–De. In summary, while hub texts somewhat depend on languages, they also achieve certain scores in non-target languages.

5.3 Score distribution

We show the scatter and box plots of COMET scores for each test case in Figure 2. The results of all the steps have low SD compared with M2M100's translations. Therefore, we found that the results of each step were also evaluated consistently with high COMET scores, regardless of test cases.

5.4 Formal analysis and implementation

Complexity class Finding the optimal hub text is an NP-hard problem. This is because verifying whether a hypothesis text $\mathbf{h}$ maximizes the evaluation score requires computation over an exponential or even infinite space, i.e., $\mathcal{V}^*$. Unlike images and audio, a hub text is represented in a discrete space; thus, it cannot be searched via gradient descent, and we need to enumerate and verify all possible candidates to find the exact solution.

Our method finds an approximate solution within a feasible computational time by narrowing the search space through hub decoding and local search, rather than finding the exact solution.

Time complexity of local search In our method, step (3), the local search, is time-consuming because it has four nested loops as described in Algorithm 1. Its time complexity is $\mathcal{O}(T|\mathbf{h}||\mathcal{V}||\mathcal{D}_{\text{tune}}|)$, where $T \in \mathbb{N}$ is the number of epochs until the solution is converged [2]. Thus, we need to evaluate $T \times |\mathbf{h}| \times |\mathcal{V}| \times |\mathcal{D}_{\text{tune}}|$ scores. Here, the loops for $\mathcal{V}$ and $\mathcal{D}_{\text{tune}}$ are highly concurrent because each iteration can be computed independently, enabling parallel computation by leveraging GPUs. The vocabulary $\mathcal{V}$ of COMET contains 250K tokens, and $\mathcal{D}_{\text{tune}}$ has 2,074 sentence pairs in En–Ja. We split the vocabulary dimension and created a mini-batch for each chunked vocabulary. In our experiments, we computed in 6,160 seconds with 8 NVIDIA RTX 6000Ada GPUs for the local search in En–Ja. Concretely, each token was replaced with one that maximizes the evaluation score in 44 seconds, and a total of 140 tokens were replaced in Algorithm 1.

6 Conclusion

We proposed a method for finding hub texts that receive unreasonably high scores regardless of references and was the first to reveal critical vulnerabilities in the neural evaluation metric COMET. Our experiments showed that a single hub text achieved higher COMET scores than M2M100's translations, even though M2M100 translated each source sentence individually, in the WMT'24 En–Ja and En–De translation tasks. These results suggest that relying on a single evaluation metric is unreliable, and the existence of hub texts further reaffirms the need for multi-metric evaluation.

[2]From our preliminary experiments, in most cases, T was less than 10.

Limitations

Scope We proposed a method for finding adversarial hub texts in the discrete text space to investigate the reliability of neural evaluation metrics for text generation tasks. This paper demonstrated the vulnerability in a case study of COMET, but our method can be applied to other neural evaluation metrics. The goal of our work is to reveal the existence of hub texts in neural evaluation metrics. In this short paper, we employed COMET with a model of `Unbabel/wmt22-comet-da`[3] for the target metric, and En↔Ja and En↔De translation tasks for the evaluation sets. Therefore, we have already conducted a comprehensive analysis across four translation directions, confirming that the hub text derived from the En–Ja setting generalizes cross-lingually, including Ja–En, En–De, and De–En. We believe *this short paper presents sufficient evidence to support our claims*, as encouraged by the ACL Reviewer Guidelines (H13)[4]. Exhaustively identifying hub texts across all existing metrics and languages is out of our scope.

Detectability Some readers may argue that hub texts could easily be filtered out. However, such a claim is akin to the *egg of Columbus*: these concerns arise only because our work has revealed and demonstrated the existence of the hubness problem. Moreover, naive COMET scores are already being used in practical scenarios, such as system filtering and ranking (see Section 5.1). This demonstrates that our identified vulnerability is already impactful under current real-world settings. Importantly, our core contribution lies in demonstrating that even a single, unnatural hub text can receive unreasonably high evaluation scores when appropriate filtering mechanisms are not in place. In addition, we propose a method for finding both natural and unnatural hub texts. While the output of step (3) in our algorithm tends to be unnatural, the output of step (2) remains fluent and readable to humans, making automatic detection significantly more difficult. This short paper presents a single, well-defined contribution: identifying a specific vulnerability in a widely used neural evaluation metric and proposing effective methods to discover adversarial hub texts.

Resoureces Our method requires expensive computational resources due to step (3), Algorithm 1, which is time-consuming. We further discuss the time complexity of the local search in Section 5.4.

Ethical Considerations

Licenses We used the publicly available benchmark datasets, WMT'23 and WMT'24 general translation tasks, and JParaCrawl v3 in accordance with these licenses. The details of these licenses are described in Appendix A.

Potential risks Abuse of our method could lead to cheating in competitions, false hypes, and other problems. To avoid these problems, it is important to evaluate using multiple metrics without relying too heavily on a single metric and to conduct a human evaluation. We hope that our method will help identify vulnerabilities in evaluation metrics, raise awareness of their reliability issues, and contribute to the development of more robust and trustworthy evaluation methods. Therefore, we publish this study to proactively address the possibility of adversarial attacks before they occur. Importantly, the risks we highlight were not introduced by our method, but are inherent risks of the evaluation metrics themselves. In this sense, our method does not amplify the risk; rather, it reveals a pre-existing "ticking time bomb" that has always been present.

Co-ordinated disclosure This study focuses on reporting the weaknesses of embedding-based neural metrics, falling under the category of coordinated disclosure. While the methods we introduce are designed to expose specific vulnerabilities in evaluation models, it is still possible for users to unintentionally input such adversarial text, even without explicitly employing our proposed techniques. This suggests that the issues we raise could realistically occur in out-of-the-box settings. Therefore, we believe our work does not fall under the definition of *coordinated disclosure* as stated in the ACL Policy on Publication Ethics[5]. Moreover, this hub text was initially disclosed on March 14, 2025[6], which predates April 25, 2025, the effective

[3] We investigated this model as it is the de facto standard model among automatic evaluation metrics for translation tasks. It is employed as the default model when no model name is specified in `comet-score` (https://github.com/Unbabel/COMET), the command-line interface of COMET.

[4] https://aclrollingreview.org/reviewerguidelines#review-issues

[5] https://www.aclweb.org/adminwiki/index.php/ACL_Policy_on_Publication_Ethics#Co-ordinated_disclosure

[6] Oral presentation at a shared task workshop. (https://sites.google.com/view/nlp2025ws-langeval/task/translation)

date of this policy. Since multiple official procedures, including patent filings, have been carried out based on this disclosure date, we believe that this hub text does not fall within the scope of the current policy.

Acknowledgements

We thank all the reviewers of this paper for their constructive and valuable feedback.

This study is inspired by the "metric hack shared task" held at the "NLP2025 Workshop: Present and Future of Natural Language Evaluation in the LLM Era", which is a workshop held in conjunction with the domestic conference "NLP2025" in Japan. This shared task aimed to identify weaknesses in existing automated evaluation metrics, and our team, "きょなら", won the competition and received the "Best Hacking Award", by identifying a vulnerability based on hubness as described in this paper. We completed this paper by expanding and generalizing the methods developed during the shared task period, and adding many experiments and analyses. We sincerely appreciate Katsuhito Sudoh (Nara Women's University), the organizer of the translation track in the shared task, as well as all the workshop organizers, including Mamoru Komachi (Hitotsubashi University), Tomoyuki Kajiwara (Ehime University), and Masato Mita (Recruit Co., Ltd.).

References

Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. 2024. Nearest neighbor normalization improves multimodal retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22571–22582, Miami, Florida, USA. Association for Computational Linguistics.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. *Preprint*, arXiv:1412.6568.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

John Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander Rush. 2023. Text embeddings reveal (almost) as much as text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12448–12460, Singapore. Association for Computational Linguistics.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There‘s no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Yimu Wang, Xiangru Jian, and Bo Xue. 2023. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10542–10567, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tingwei Zhang, Fnu Suya, Rishi Jha, Collin Zhang, and Vitaly Shmatikov. 2025. Adversarial hubness in multi-modal retrieval. *Preprint*, arXiv:2412.14113.

A Licenses

Datasets We used the WMT'23 Ja↔En and De↔En, and WMT'24 En–Ja and En–De general translation tasks, released under the policy[7,8]: "The data released for the WMT General MT task can be freely used for research purposes". To train the inversion model, we used JParaCrawl v3, licensed by Nippon Telegraph and Telephone Corporation (NTT) for research use only as described in `http://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/`.

Models We used `Unbabel/wmt22-comet-da` for the COMET metric and `google/mt5-base` for the inversion model, released under the Apache-2.0 license. We also used MIT-licensed `facebook/m2m100_418M` for comparison purposes[9].

B Dataset Statistics

Table 6 lists the statistics of the datasets we used.

[7] `https://www2.statmt.org/wmt23/translation-task.html`

[8] `https://www2.statmt.org/wmt24/translation-task.html`

[9] We used M2M100 because it has been cited by over 900 papers and is suitable for the baseline translation model.

Dataset	Purpose	Notation	#examples
WMT'23 En–Ja test set	Tuning (Development)	$\mathcal{D}_{\text{tune}}$	2,074
WMT'24 En–Ja test set	Test (Evaluation)	$\mathcal{D}$	997
WMT'23 En–De test set	Tuning (Development)	$\mathcal{D}_{\text{tune}}$	557
WMT'24 En–De test set	Test (Evaluation)	$\mathcal{D}$	997
JParaCrawl v3 (Ja only)	Training of inversion model	$\mathcal{D}_{\text{mono}}$	1,000,000
En–De parallel corpus from CommonCrawl (De only)	Training of inversion model	$\mathcal{D}_{\text{mono}}$	2,399,123
WMT'23 Ja–En test set	Test (Table 5)	$\mathcal{D}$	1,992
WMT'23 De–En test set	Test (Table 5)	$\mathcal{D}$	549

Table 6: Statistics of datasets we used