

SEScore2: Retrieval Augmented Pretraining for Text Generation Evaluation

Wenda Xu[¶], Xian Qian[†], Mingxuan Wang[†], Lei Li[¶], William Yang Wang[¶]

[†]ByteDance, [¶]UC Santa Barbara

{wendaxu, leili, william}@cs.ucsb.edu

{wangmingxuan.89, qian.xian}@bytedance.com

Abstract

Is it possible to leverage large scale raw and raw parallel corpora to build a general learned metric? Existing learned metrics have gaps to human judgements, are model-dependent or are limited to the domains or tasks where human ratings are available. In this paper, we propose SEScore2, a model-based metric pretrained over million-scale synthetic dataset constructed by our novel retrieval augmented data synthesis pipeline. SEScore2 achieves high correlation to human judgements without any human rating supervisions. Importantly, our unsupervised SEScore2 can outperform supervised metrics, which are trained on the News human ratings, at the TED domain. We evaluate SEScore2 over four text generation tasks across three languages. SEScore2 outperforms all prior unsupervised evaluation metrics in machine translation, speech translation, data-to-text and dialogue generation, with average Kendall improvements 0.158. SEScore2 even outperforms SOTA supervised BLEURT at data-to-text, dialogue generation and overall correlation ¹.

1 Introduction

In recent years, researchers made significant progresses in text generation: translation (Birch, 2021), structured data-to-text (Gardent et al., 2017), dialogue generation (Vinyals and Le, 2015), summarization (Chopra et al., 2016) and image captioning (Fang et al., 2014). A well-developed metric can efficiently indicate the model performance (Celikyilmaz et al., 2020) and guide the text generation (Unanue et al., 2021; Freitag et al., 2022a). Without the proper automatic metrics, many researches can be hindered and even misled to wrong directions (Callison-Burch et al., 2006).

Depending on the inputs to the metrics, we can categorize evaluation metrics into source-based, hybrid-based and reference-based metrics. The

source-based metrics directly estimate the quality of the text through source and are useful when reference is noisy or unavailable (Louis and Nenkova, 2013; Kepler et al., 2019). However, they are often sub-optimal due to the gap between source and reference and easy to explore spurious correlations (Durmus et al., 2022). Hybrid metric COMET (Rei et al., 2020) computes quality score from both source and reference. In practice, many text generation tasks have non-text source (e.g audio and image), further adding the complexity to the pipeline. Ideally, if the generated texts can be paired with the high-quality references, the proximity between references and candidate outputs should reflect the quality of text generation. In this work, we focus on building a general sentence-level, reference-based metric, invariant to the source modalities.

The rule-based metrics (e.g n-gram matching BLEU (Papineni et al., 2002), chrF (Popović, 2015)) and distance-based (e.g. TER (Snover et al., 2006)) have predominated text generation evaluations because they are fast and invariant to different domains. However, they are bounded by the surface form differences and unable to capture semantics and long distant dependencies (Zhang et al., 2019). Recently, many researches have encoded learned components into the evaluation process. **Supervised learned metrics** attempt to optimize directly from the human ratings (Rei et al., 2020; Sellam et al., 2020). **Unsupervised learned metrics** obtain training objectives other than human ratings: MLM pretraining (Zhang et al., 2019) or sequence-to-sequence pretraining (Thompson and Post, 2020; Yuan et al., 2021).

Our goal of this paper is to devise a **reference-based** automatic evaluation metric that can be learned from raw or supervised corpora without using human-annotated reference-candidate text pairs. Ideally, a learned metric for language X should be generalized to evaluate different domains and different NLG tasks. To achieve such goals,

¹<https://github.com/xu1998hz/SEScore2>

our data synthesis pipeline should contain realistic and diverse set of errors and produce a large-scale of synthetic data, to cover model mistakes at different domains and tasks. We develop a novel retrieval augmented data synthesis pipeline which models surface form differences between neighboring raw text pairs, yielding a proposal of non-overlapping, plausible text transformations. A random subset of the proposed operations are applied to the raw text to simulate co-occurrences of errors in a model output. This data construction process is independent on any generative models and can build five million synthetic dataset within ten minutes. Inspired by the severity label of the human grading process (Freitag et al., 2021a), we leverage a large scale raw corpus and cross-lingual MLM model (Conneau et al., 2019) to estimate severity measure of each transformation in the proposal before applying them to the raw text. Our contributions of this paper are as follows:

- We leverage raw and raw parallel texts to pre-train a learned metric without human ratings;
- We propose a novel retrieval augmented data synthesis pipeline to construct hard negative samples and empirically demonstrate its superior performance compared to the random token transformations;
- We annotate additional human ratings for WMT21 German-to-English testing set following MQM human annotation procedure;
- Experiments show that SEScore2 is effective in a diverse set of text generation tasks: machine translation, speech translation, data-to-text and dialogue generation and outperforms all unsupervised metrics significantly. It even outperforms SOTA supervised metric BLEURT in overall Kendall correlation.

2 Related Work

The supervised learned metrics (Rei et al., 2020; Sellam et al., 2020) can fit tightly to the human rating distribution. However, they may have poor generalization to unseen domains and tasks (Freitag et al., 2021b). Unsupervised learned metrics attempt to obtain training objectives other than human ratings (Zhang et al., 2019; Zhao et al., 2019; Thompson and Post, 2020; Yuan et al., 2021). However, as pointed out by (Freitag et al., 2021a), they

are limited on the error types that they can evaluate (e.g accuracy) and can not go beyond to (e.g fluency or style). Some recent studies attempt to mitigate this issue by generating synthetic data via paraphrasing and perturbations (Sellam et al., 2020; Kryscinski et al., 2020; Gao et al., 2021). To further derive the fine-grained pretraining signals, SEScore (Xu et al., 2022) leverages language models to generate multiple error types in one segment and estimate each error’s severity level. However, model-dependent data synthesis can intrinsically introduce model bias and limit the diversity of data samples. SEScore2 develops a novel retrieval augmented data synthesis technique, which simulates diverse and realistic model mistakes purely on raw corpus, independent on any generative model.

The WMT shared metrics competition (Ma et al., 2018, 2019; Mathur et al., 2020) is the main submission entry for evaluation metrics. However, their crowd-worker evaluation using direct assessment (DA) has been questioned (Läubli et al., 2020). Mathur et al. (2020); Freitag et al. (2021a) find that crowd-workers fail to discriminate human and machine outputs. Freitag et al. (2021a) improves human ratings by using Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) with language experts. Each annotated error can be categorized into multiple types and is associated with different levels of severity, such as major and minor. Taking the inspiration from (Freitag et al., 2021a), we obtain our pretraining signals aligning with human grading process. Recent findings in WMT21 (Freitag et al., 2021b) suggest that the performance of metrics largely varies depending on the underlying domains. We strengthen this argument by demonstrating our unsupervised SEScore2 can outperform supervised metrics, which trained on News human ratings, at TED domain.

3 Preliminaries

Problem Definition The problem of the learned evaluation metric contains training and inference components. Ideally, given a set of references y , candidate outputs x and quality scores s , we can train a supervised learned metric, M_s , from such triple set (x, y, s) so that M_s can produce a quality score s^* for unseen pair (x^*, y^*) during inference. Due to the scarcity of human ratings, triples (x, y, s) are unavailable for most of tasks. Therefore, for the unsupervised learned metric M_u , it has to leverage either raw data or raw parallel data

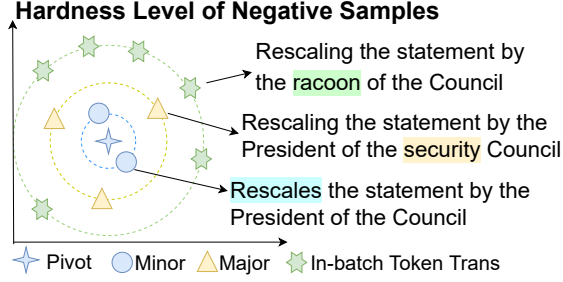


Figure 1: 4-point star represents the pivot sentence. Pivot sentence: *Rescaling the statement by the President of the Council*. Circles and triangles represent the sentences with minor and major mistakes, respectively. Both triangles and circles are hard negatives. 6-point stars are easy negatives produced by the random token transformations. Circles that are inner indicates the negative samples are harder.

other than (x, y, s) to obtain the training signals, Q . During inference, M_u , which is trained from objective Q , can be used to output a score s^* for unseen pair of (x^*, y^*) .

Hard Negative Construction Constructing hard negative samples is the most essential step in the embedding learning. One common practice is to apply in-batch token insertions or replacements to the pivot sentence (Fu et al., 2022). However, as shown in Figure 1, in-batch text transformations mostly produce the negative samples that are syntactically, semantically incorrect (Chomsky, 1956) or both. We consider those constructed samples as easy negatives, as they largely deviate from general model mistakes (Freitag et al., 2021a). In practice, the majority of the model-generated errors can be semantically sound or syntactically correct, see triangle and circle examples in Figure 1. Following (Freitag et al., 2021a), if the model error semantically alters the meaning of the sentence, we label them as major and minor otherwise. In both cases, they are considered as hard negatives because an ideal metric needs to capture the subtle changes between model outputs and references, in considering severity levels. This motivates us to propose a retrieval augmented data synthesis approach, which scales the diversity of synthetic texts along with the size of the raw corpus, including both major and minor error types, all three cycles in Figure 1.

4 The SEScore2 Approach

4.1 Overview

SEScore2 is an unsupervised learned evaluation metric, initialized from the pretrained language model (e.g BERT). From a raw text corpus, we extract text pair (x_0, x_j) , such that they are the k -nearest neighbor of one another. From the surface form difference of x_0 and x_j , along with the random word drops, we derive a proposal, P for stratified error synthesis (Xu et al., 2022), $P = (P_1, P_2, \dots, P_n)$, such that x_n can be produced through n recursive text transformations from the proposed operations (eq.1).

$$\mathbf{x}_i = \begin{cases} \mathbf{x}_0, & \text{if } i = 0 \\ P_i(\mathbf{x}_{i-1}), & 0 < i \leq n \end{cases} \quad (1)$$

We construct a severity scoring function, S_i , which takes in raw text x_0 and current proposal P_i . At each step, we estimated the severity score of the current proposal, $S_i(x_0, P_i)$. We randomly select k proposals from the existing n proposals. After k steps, we cumulatively yield the overall quality score for the transformed text x_k , with the scoring label $s_k = \sum S_i(x_0, P_i)$. Please see Figure 2 for a concrete example. This stratified error synthesis approach can yield synthetic training triples (x_0, x_k, s_k) to train our unsupervised quality prediction model. During inference, our quality prediction model can produce scores for all unseen reference and candidate pairs. Details of quality prediction model can be found in Appendix B.

4.2 Pretraining on Synthetic Data

The key aspect of SEScore2 is a continue pretraining technique (Gururangan et al., 2020) that learns NLG evaluation objective during pretraining and can be inferenced directly for any text generation evaluation without fine-tuning. Ideally, the pretraining data distribution should resemble closely to the final NLG evaluation. For example, the candidate sentence outputs can come from real NLG model outputs. However, this can automatically introduce model output biases during the learning process.

To pretrain a general learned metric, we model our pretraining data construction into four constraints: 1) The pretraining data should come from a large diverse raw corpus. Therefore, SEScore2 can be generalized for different text domains and NLG tasks. 2) The synthetic data construction should be fast and independent on any text generation models. 3) The synthetic data should contain

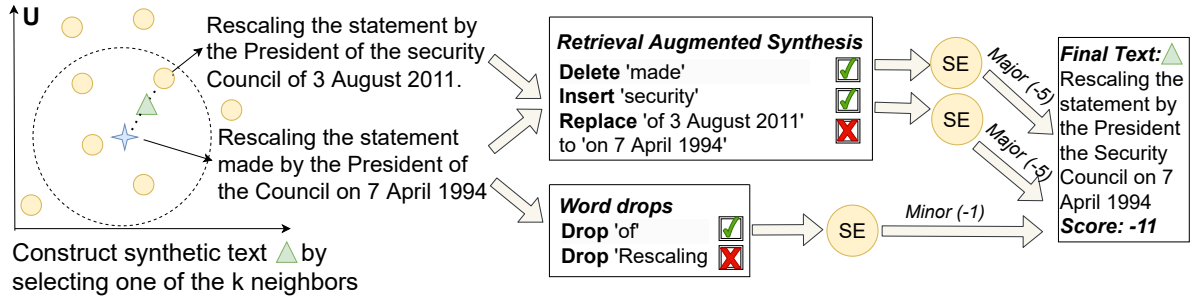


Figure 2: Retrieval Augmented Data Synthesis: we denote raw pivot text, selected neighbor, and synthesized text as 4-point star, circle and triangle respectively. We randomly select a subset of proposed transformations (ticks) and estimate severity measures (SE) on them. Final score sums the individual severity measures. Final text, raw pivot text and estimated score will be used to train our quality prediction model.

realistic and diverse model mistakes, in order to incorporate different aspects of text generation errors, such as accuracy and fluency. 4) The pretraining signals should resemble how humans grade the model outputs, namely to estimate the severity levels of each mistake (Freitag et al., 2021a).

4.2.1 Retrieval Augmented Data Synthesis

Starting from a raw text sentence x_0 , SEScore2 first map x_0 into the embedding space e_0 and find the k nearest neighbors to e_0 through the index table search. Detailed index table construction can be found in Appendix C. We randomly pick one of those k nearest neighbors e_j and return its text form x_j . We use edit distance operations to model the surface form differences between x_0 and x_j . In another words, x_0 can go through a chain of edit operations P , $P = (Insert_{a \rightarrow b}, Replace_{c \rightarrow d}, Delete_{e \rightarrow f}, \dots)$, where subscripts are modified text spans in sentence x_0 , to transform into x_j . These edit operations, augmented with random word drops at unmodified text locations, naturally yield a set of non-overlapping insert, replace, delete operations, with respective locations to perform. Therefore, given a raw text $x_0 = (t_1, \dots, t_n)$ with its neighbor x_j in the raw corpus, t_i indicates i th token in x_0 , SEScore2 can obtain proposed text transformations, P . To construct the synthesized text x_k , we randomly select a subset of k text transformations from the proposal P and recursively apply transformation to x_0 . In Figure 2, we demonstrate the data synthesis process from sentence pair (x_0, x_j) and show an example of synthetic text x_k that is constructed from the proposal. The entire retrieval augmented data construction is in CPU-operations and independent on any generation model. **This process can produce**

five million sentences within 10 minutes.

4.2.2 Synthesizing Diverse Model Errors

Synthesizing realistic and diverse hard negative pretraining data are the core challenges to learn an evaluation metric. As the recent study showed (Freitag et al., 2021b), many errors in the model outputs tend to be fluent, however, semantically deviated from the reference. To synthesize such errors, we control the text pair proximity by setting a margin criterion when retrieving the k nearest neighbors (Schwenk et al., 2021). Based on our preliminary study, margin criterion ($k = 1.06$) can retrieve sentences with similar text structure or semantics. Detailed margin criterion formulation can be found in the Appendix D. In figure 2, we demonstrate that our retrieval augmented data construction can synthesize errors which contextually sound but semantically deviate from the reference. To maximize the diverse error types during synthesis, we add two levels of flexibility during the proposal generation. First, all sentences $\{x_j\}$ that above the margin to the pivot sentence x_0 have the same probability to be selected to construct the proposal (All circles in the loop have the equal chances to be selected in Figure 2). Therefore, a diverse set of lexical, syntactic and semantic dissimilarities will be considered during retrieval augmented construction. Second, we randomly select 1 to 5 transformations from the proposal so that synthesized text will be stratified and contain a diverse number of errors. For sentences that contain more than five errors will be labelled as the catastrophic errors with the lowest rating -25 according to the human evaluation (Freitag et al., 2021a). In Appendix H, we demonstrate that our data synthesis pipeline can cover a diverse set of error types. One last

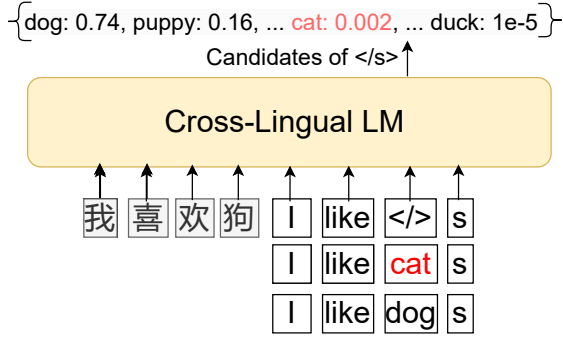


Figure 3: Source Chinese text means 'I like dogs'. Our retrieval augmented data synthesis replaces 'dog' with 'cat'. Therefore, we replace 'cat' with special token '</s>' and estimate the probability of recovering '</s>' to 'cat' given the context of source and rest target tokens. If $p(cat|contexts) \geq \gamma$, it is a minor mistake. If otherwise, it is a severe mistake.

difficulty is to synthesize positive samples of the pivot, since no ground truth is available. Inspired by (Gao et al., 2021), we leverage dropout function to simulate paraphrasing embeddings by feeding the pivot twice (score: 0). Lastly, in-batch negatives are included for the similarity lower bound (score: -50)².

4.2.3 Pretraining Signals

Inspired by the human grading process (Freitag et al., 2021a), we consider two levels of severity measures: major and minor, for each error in the candidate outputs. An error is major if it alters the core meaning of the sentence. To model this definition, we develop two severity measure techniques to estimate different operations separately. For insert and replace operations, we adopt a cross-lingual MLM model to perform the severity estimation, such as XLM (CONNEAU and Lample, 2019). We first mask the perturbed locations of text x_0 to obtain $x_{mask,i}$ where x_i is produced by a single text transformation P_i on x_0 . To formulate this estimation, our severity estimation model will take in three inputs: source tokens $s = (s_0, \dots, s_m)$; masked target tokens, $x_{mask,i} = (x_0, \dots, x_n)$, where perturbed locations are masked; perturbed tokens $m = (m_0, \dots, m_j)$. Then, XLM takes in concatenated source text s with $x_{mask,i}$ and estimate the probability p to recover span m from masks. The intuition is that XLM with TLM objectives can model co-occurrences of source, target tokens and word alignments between parallel

sentences. If our retrieval augmented operation P_i alter the original sentence meaning, their corresponding mask location will have low probabilities to recover the perturbed tokens. By setting a threshold γ , we assign minor label to the transformation T_i if $p_{mask} \geq \gamma$ and severe if $p_{mask} < \gamma$. A concrete example can be found in Figure 3. For delete operation, our main assumption is that delete operation creates minor error if the deleted tokens do not contain meaningful semantics in the sentence. In this case, we use token importance weight to approximate the importance of semantics that each token conveys. We adopt TF-IDF weights to estimate the importance weights. By setting a threshold λ , if the importance weights w of the modified span, $w \leq \lambda$, the transformation T_i is minor and if $w > \lambda$, error is severe. In the end, pretraining signal for each segment is the sum of individual severity measures³.

5 Experiments

To verify the generalizability and utility of the SEScore2, we investigate the following questions: 1) Can SEScore2 be generalize to multiple domains of the same task? 2) Can SEScore2's language checkpoint X be used to evaluate all languages Y to X's outputs? 3) Can SEScore2's language checkpoint X be used to evaluate all different text generation tasks on language X? 4) How to interpret SEScore2? Corresponding to the aforementioned evaluation aspects, 1) we test SEScore2 over two different domains (News and TED) of machine translation task. 2) we test SEScore2's English checkpoint over multiple Y-to-English directions to assess its performance on arbitrary language to English's performance. 3) we test SEScore2's English checkpoint over diverse set of text generation tasks: Machine Translation, Speech Translation, Data-to-Text and Dialogue Generation. 4) we decompose SEScore2 into multiple evaluation dimensions and interpret overall score from each dimension. Moreover, we conduct comprehensive experiments for each component of SEScore2 and analyze the leading factors contributing to the final result. We conduct William's pair-wise significance test (Graham and Baldwin, 2014) to highlight the significant improvements.

²We set score -50 to distinguish easy and hard negatives (-1 to -25). However, -50 is not a sensitive hyperparameter.

³ $\lambda = 1$ and $\gamma = 0.1$ for all three languages.

Language	Index Table		Pretraining Data	
	News	Wikipedia	Pivot	Retrieved
English	20M	20M	5M	13.5M
German	4.5M	16M	4.5M	13.2M
Japanese	18M	12M	5M	13.3M

Table 1: Statistics for Index table and pretraining data.

5.1 Pretraining Step

5.1.1 Pretraining Data

For Chinese-to-English, we collect 20M sentence pairs from UN Parallel (Ziems et al., 2016), News Commentary (Tiedemann, 2012) and CWMT corpus (Barrault et al., 2019). For English-to-German, we collect 4.5M sentence pairs from Europarl (Koehn, 2005), Common Crawl (Kúdela et al., 2017) and News Commentary. For English-to-Japanese, we collect 18M sentence pairs from News Complimentary, WikiMatrix (Schwenk et al., 2021) and En-Ja subtitle corpus (Pryzant et al., 2018). We randomly sampled 5M, 4.5M and 5M sentence pairs for Zh-En, En-De and En-Ja respectively. We use each target language sentence as a pivot to retrieve 128 nearest neighbors to build index table and use parallel sentences to compute severity measures. We train separate checkpoint for each language direction and we use the final English checkpoint to evaluate SEScore2 in different text generation tasks. To ensure our synthetic data construction can cover general text domains, our index table includes collected News domain target sentences and raw corpus from Wikipedia dumps. Specific statistics are included in Table 1.

5.1.2 Scoring Model

To ensure fair comparisons to the SOTA model COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) at Machine Translation, we benchmark two backbones: RemBert (Chung et al., 2020b) (used by BLEURT) and XLM-Roberta-large (Conneau et al., 2019) (used by COMET). Detailed analysis is included at Section E. To ensure the fair model size comparisons, we use the XLM-Roberta-large backbone for Data-to-text and Dialogue generation. We use Adam optimizer and set bath size, learning rate and drop out rate to be 256, 3e-5 and 0.1 respectively. We use the mean squared error to train the metric model. All checkpoints from rembert and xlm-roberta-large is trained for 15,000 and 30,000 iterations respectively.

LP	News			TED		
	#H	#Sys	#Sents	#H	#Sys	#Sents
Zh→En	2	13	650	1	13	529
En→De	4	13	527	1	14	529
De→En	1	9	100	-	-	-

Table 2: Human annotation statistics for Machine Translation (MT) task. #H refers to the number of humans, #Sys refers to the number of MT systems and #Sents refers to the number of annotated samples per system.

5.2 Baseline Model

For all text generation tasks, we include 1) three n-gram, distance-based baseline metrics: BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and TER (Snover et al., 2006); 2) four best performed learned metrics without human ratings: PRISM (Thompson and Post, 2020), BARTScore (Yuan et al., 2021), BERTScore (Zhang et al., 2019) and SEScore (Xu et al., 2022); and 3) two SOTA supervised learned metrics: COMET⁴ and BLEURT.

5.3 Evaluation Procedure

For all the text generation tasks, we compute segment-level Kendall correlation between metric outputs and human scores.

Machine Translation For En-De and Zh-En, we used WMT21 News and WMT21 TED’s human annotations. We also hired 3 professional linguists to annotate 1000 testing samples from WMT News De-En direction using MQM evaluation procedures (Freitag et al., 2021a). Detailed statistics are included in Table 2 and MQM annotation procedure is discussed in Appendix.

Dialogue Generation BAGEL benchmark contains target utterance generation for spoken dialogue systems. This benchmark contains 202 model outputs. Each sample is annotated in the aspect of naturalness, informativeness and quality.

Data-to-Text Generation WebNLG contains 17 participating models and each contains 177 model outputs. Each sample is annotated by six aspects: correctness, data coverage, fluency, relevance and text structure.

Speech-to-Text We use IWSLT 22 English-to-Japanese (En-Ja) human annotations. The bench-

⁴Since COMET is a source-reference-based approach only applicable to translation tasks, we only used to generate results for machine and speech translation

		MT(Zh→En)	MT(En→De)	MT(De→En)	S2T(En→Ja)	D2T	Dialogue	Overall
With	BLEURT	0.291	0.252	0.266	0.463	0.168	0.229	0.278
	COMET(DA)	0.290	0.249	0.250	0.405	-	-	-
Without Supervision	TER	0.173	0.115	-0.046	-0.082	-0.090	-0.087	-0.003
	BLEU	0.134	0.098	0.068	0.202	0.084	0.109	0.116
	ChrF	0.158	0.130	0.074	0.240	0.094	0.108	0.134
	BARTScore	0.208	0.042	0.047	-0.123	0.113	0.203	0.082
	BERTScore	0.248	0.179	0.205	0.213	0.154	0.171	0.195
	PRISM	0.240	0.215	0.174	0.198	0.163	0.217	0.201
	SEScore	0.281	0.226	0.249	0.361	0.155	0.192	0.244
	SEScore2	0.310	0.243	0.250	0.458	0.182	0.233	0.279

Table 3: Segment-level Kendall correlation on En-De, De-En and Zh-En for WMT21, En-Ja for IWSLT22, WebNLG20 data-to-text and BAGEL dialogue generation. SEScore2 significantly outperforms all unsupervised metrics in all tasks and BLEURT in D2T and dialogue generation, with p values < 0.05.

mark contains four participating systems and each system contains 118 outputs. All human annotations are done using JTF MQM variant (JTF, 2018).

5.4 Overall Performance

In Table 3, we demonstrate metrics’ overall performance in machine translation, speech translation, data-to-text and dialogue generation. SEScore2 outperforms all rule-based metrics significantly in the overall Kendall correlation, with the average absolute Kendall improvement of 0.197. SEScore2 outperforms all unsupervised learned metrics significantly in all four tasks and three MT translation directions. More importantly, SEScore2 outperforms the supervised BLEURT in two of the four text generation tasks and achieve higher Kendall correlation overall across four tasks. We will discuss each task in details in Section 5.5.

5.5 Performance across different tasks

For Machine Translation, SEScore2 outperforms all unsupervised metrics significantly across all three language directions. SEScore2 outperforms both supervised metrics COMET and BLEURT in Zh-En. Furthermore, SEScore2 achieves comparable performance to COMET and close performance to BLEURT in En-De and De-En. This is a significant result as all three language directions are presented at training set of BLEURT and COMET. At speech translation, SEScore2 outperforms all unsupervised metrics significantly and leads COMET by a large margin. One possible reason to explain the large improvement between COMET and SEScore2 is that English to Japanese human ratings are not presented at COMET’s supervision data and this further exposes vulnerabilities of the supervised metrics at unknown domain

or language direction. Lastly, SEScore2 outperforms all supervised and unsupervised metrics at data-to-text and dialogue generation. Compared to BLEURT, which is supervised on translation human rating, SEScore2 can achieve superior generalization capability in non-translation tasks, such as data-to-text and dialogue generation.

5.6 Performance over Y→X Translations

For machine translation task, we further investigate SEScore2’s generalization capabilities over different languages to X translation outputs. In particular, we compared SEScore2 at Zh→En and De→En directions. From Zh→En, SEScore2 outperforms all rule-based and unsupervised learned metrics significantly. The pattern is consistent in De→En, where SEScore2 outperforms all unsupervised metrics significantly except SEScore. Compared to the supervised learned metrics, SEScore2 outperforms both BLEURT and COMET at Zh→En. SEScore2 achieves the comparable performance to the COMET and 1.6% Kendall correlation gap to BLEURT. This is a significant result as both Zh→En and De→En were included in the supervised training dataset for COMET and BLEURT and SEScore2 can achieve the close-to-supervised performance. We will discuss the domain influence of language direction in Section 5.7.

5.7 Performance across different domains

We investigate the domain influence on the evaluation metrics when shifting testing set from News to TED. As shown in Table 4, all metrics have lower Kendall correlations in TED compared to those in News. We conjectured that the cause is due to the domain differences in two testing suites. Unlike News, TED contains sentences with infor-

	Model Name	Machine Translation (WMT21)			
		News	TED	Overall	Δ
With	BLEURT	0.305	0.243	0.274	0.062
	COMET(DA)	0.300	0.240	0.270	0.060
W.o Supervision	TER	0.154	0.134	0.144	0.020
	BLEU	0.130	0.103	0.117	0.027
	ChrF	0.158	0.135	0.147	0.023
	BARTScore	0.140	0.111	0.126	0.029
	BERTScore	0.232	0.194	0.213	0.038
	PRISM	0.239	0.216	0.228	0.023
	SEScore	0.273	0.235	0.254	0.038
	SEScore2	0.287	0.265	0.276	0.022

Table 4: Segment-level Kendall correlation for WMT21 (En-De and Zh-En) News and TED Testing sets. Δ indicates the absolute correlation difference between News and TED. Overall indicates the metrics’ average performance of News and TED domains.

mal and disfluent language styles (Freitag et al., 2021b). The supervised learned metrics have the largest performance gap when shifting domain from News to TED. The root reason is that the entire supervised human rating data is from News domain only. Although the rule-based metrics (TER, BLEU and ChrF) have relatively lower overall correlations, their correlation is less influenced by the domain shift, with average 0.023 Kendall correlation difference. Unsupervised learned metrics can be less influenced by domain shift compared to the supervised metrics. However, they still have more Kendall correlation drops compared to the rule-based metrics, with average Kendall correlation 0.032. Most importantly, we observed that SEScore2 achieves the highest overall Kendall correlation and achieve the lowest gap (0.022) when shifting between domains. In Section 5.8.1, we demonstrate that SEScore2 can take advantage of large data scale and improve its performance on TED domain while scaling up the data. We include SEScore2’s performance in other domains of WMT22 in the Appendix Table 6 and Table 7.

5.8 Quantitative Analysis

To validate the ideas in SEScore2, we analyze the effects of pretraining data quantity, supervision from human ratings, severity measures and data constructions. We include effects of model initialization in Appendix E and error analysis in the Appendix I.

5.8.1 Law of the Data Scaling

We study the scaling effects on SEScore2’s pretraining performance, by testing checkpoints trained at

Metric Name	Machine Translation WMT21			
	News	TED	Overall	Δ
COMET	0.300	0.240	0.270	0.060
BLEURT	0.305	0.243	0.274	0.062
FT SEScore2	0.312	0.229	0.271	0.083
SEScore2	0.287	0.265	0.276	0.022

Table 5: Segment-level Kendall correlation under the unsupervised and supervised SEScore2 with supervised COMET and BLEURT at WMT21 News and TED testing sets. Overall measures the overall correlation between two domains and Δ indicates absolute correlation difference between two domains. FT stands for fine-tuning.

0.5M, 1M, 2M and 4M training samples. For both Zh-En and En-De across two domains, we observe the sharp performance improvements at first 1M pretraining data. We observe that larger pretraining data quantity can overall lead to higher human correlations for both language directions. For Zh-En, we obtain 2.5% and 1.8% Kendall correlation improvements at 2M and 3M data points respectively. For En-De, we obtain 2.5% and 1.1% Kendall correlation improvements at 2M and 3M data points respectively. The performance starts to saturated from 3M to 4M synthetic data (around 0.5% improvements in both language directions). This result suggests that data scale around 1M can train a metric with competitive performance and larger pretraining data can gradually improve pretrained metric to fit into a general domain.

5.8.2 Danger of Fine-tuning

To address the question "Can further fine-tuning leads to better performance of SEScore2?", we fine-tune SEScore2 over existing 350K English and 59K German WMT17-19 News domain human rating data⁵. In Table 5, supervised SEScore2 gains 8.7% improvements over unsupervised version at News. Moreover, we observe that supervised SEScore2 can outperform both supervised BLEURT and COMET at News domain with 4% and 2.3% Kendall improvements respectively. However, with this improvements in News, we also observe that supervised SEScore2 has 13.6% correlation drop at TED testing set, causing a larger correlation gap (0.083) between News and TED domain. This observation verifies our assumption

⁵Identical to COMET, We use the DA human rating data from WMT17-19 for supervised SEScore2. BLEURT uses WMT15-19 DA results as its training dataset

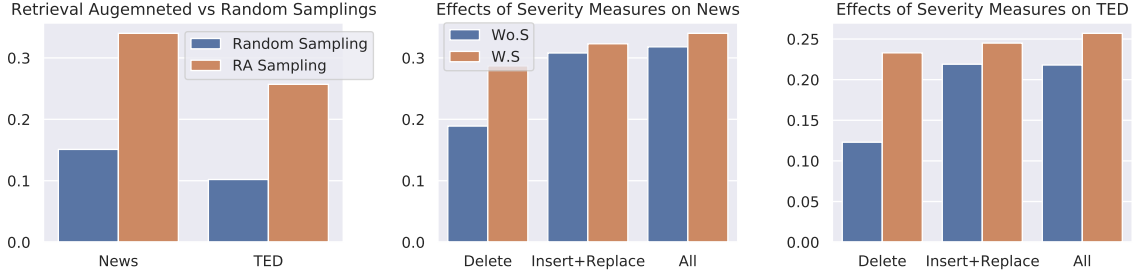


Figure 4: Left figure indicates the comparisons between SEScore2 trained from retrieval augmented data construction and random token transformations. Middle and right figure indicate individual operations contribute to final SEScore2 and effects of severity measures in News and TED domains.

that fine-tuning over domain-specific human rating data can fit learned metric tightly to the trained domain distribution. At the same time, however, this may hurt generalization capability of metrics at different domains. Unsupervised SEScore2 achieves the highest overall Kendall correlation across two domains.

5.8.3 Effects of Different Negative Sampling

In Figure 4, we demonstrate the superior performance of retrieval augmented (RA) data construction compared to the random token transformations. Our observation is that most of random in-batch tokens have low co-occurrence probabilities with their contexts. The sentence embeddings from those text transformations can be easily distinguished from the pivot embedding, by the pre-trained language model (Conneau et al., 2019). Therefore, further pretraining on negative samples with random token transformations does not lead to significant correlation improvements. We empirically demonstrate that RA data construction improves random insert/replace by 114% in the News domain and 114% in the TED domain.

5.8.4 Retrieval Augmented Data Synthesis

To evaluate the performance of each component at our retrieval augmented data synthesis, we separately trained checkpoints with synthetic data that 1) contains delete operation only; 2) contains insert and replace operations according to our scheduled locations; 3) contains all operations with scheduled positions. To exclude the effects from the severity measures, we do not assign severity measure for each error and instead label each sentence with the number of errors it contains. In Figure 4, we observe that our RA insert/replace contribute the most of the human correlations, 0.308 at News and 0.219 at TED. This suggests that our scheduled positions

to insert and replace are important to construct realistic synthetic sentences and learn meaningful embeddings. Despite the simple scheme, delete-only construction can achieve competitive performance, with Kendall correlation 0.189 and 0.123 in News and TED respectively. By combining all operations, the aggregated effect can further improve the Kendall correlation 3.2% at News.

5.8.5 Effects of Severity Measures

In Figure 4, we empirically verify two of our severity measures: 1) IDF-based 2) XLM-based approaches. Our IDF-based severity measures on delete operation can improve 51.9% Kendall correlation at News and 81.3% at TED. Our XLM-based severity measures on insert and delete can improve 4.9% at News and 11.9% at TED. Lastly, the joint effect of two severity measures can improve SEScore2 without severity measures by 6.92% at News and 17.9% at TED.

6 Conclusion

We propose SEScore2, a reference-based text generation evaluation metric. Without human rating data, SEScore2 outperforms all unsupervised evaluation metrics in four text generation task across three languages. We empirically demonstrate that compared to SOTA supervised metrics, SEScore2 have superior or competitive performance over different domains, language directions and NLG tasks. Our experimental results demonstrate that our scalable, retrieval augmented data synthesis can efficiently generate diverse and realistic error types, with human-aligned severity labels, achieving high correlation to the human judgements.

References

- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Alexandra Birch. 2021. [Neural machine translation 2020, by philipp koehn, cambridge, cambridge university press, isbn 978-1-108-49732-9, pages 393](#). *Natural Language Engineering*, 27(3):377 – 378.
- Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *EACL*.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. [Evaluation of text generation: A survey](#).
- Noam Chomsky. 1956. Three models for the description of language. *IRE Trans. Inf. Theory*, 2:113–124.
- Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. [Abstractive sentence summarization with attentive recurrent neural networks](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.
- Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020a. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.
- Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020b. [Rethinking embedding coupling in pre-trained language models](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Unsupervised cross-lingual representation learning at scale](#).
- Alexis CONNEAU and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. [Spurious correlations in reference-free evaluation of text generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics.
- Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. [From captions to visual concepts and back](#).
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. [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, Ricardo Rei, Nitika Mathur, Chiklu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André Martins. 2022b. [Results of wmt22 metrics shared task: Stop using bleu - neural metrics are better and more robust](#). In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi.
- Markus Freitag, Ricardo Rei, Nitika Mathur, Chiklu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. [Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.
- Zhiyi Fu, Wangchunshu Zhou, Jingjing Xu, Hao Zhou, and Lei Li. 2022. [Contextual representation learning beyond masked language modeling](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2701–2714, Dublin, Ireland. Association for Computational Linguistics.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Yvette Graham and Timothy Baldwin. 2014. [Testing for significance of increased correlation with human judgment](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don't stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Japan Translation Federation JTF. 2018. *JTF Translation Quality Evaluation Guidelines*.
- Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. [OpenKiwi: An open source framework for quality estimation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.
- Philipp Koehn. 2005. [Europarl: A parallel corpus for statistical machine translation](#). In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. [Evaluating the factual consistency of abstractive text summarization](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Jakub Kúdela, Irena Holubová, and Ondřej Bojar. 2017. [Extracting parallel paragraphs from common crawl](#). *The Prague Bulletin of Mathematical Linguistics*, 107(1):39–56.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Mqm: Un marc per declarar i descriure mètriques de qualitat de la traducció. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, (12):455–463.
- Annie Louis and Ani Nenkova. 2013. [Automatically assessing machine summary content without a gold standard](#). *Computational Linguistics*, 39(2):267–300.
- Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. [A set of recommendations for assessing human-machine parity in language translation](#). *Journal of Artificial Intelligence Research*, 67.
- Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. [Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.
- Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. [Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.
- Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. [Results of the WMT20 metrics shared task](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.
- 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*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.
- Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. [JESC: Japanese-English subtitle corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- 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.
- Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. [WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.
- Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. [RUSE: Regressor using sentence embeddings for automatic machine translation evaluation](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A study of translation edit rate with targeted human annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Brian Thompson and Matt Post. 2020. [Automatic machine translation evaluation in many languages via zero-shot paraphrasing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Inigo Jauregi Unanue, Jacob Parnell, and Massimo Piccardi. 2021. Berttune: Fine-tuning neural machine translation with bertscore. In *ACL*.
- Oriol Vinyals and Quoc Le. 2015. [A neural conversational model](#).
- Wenda Xu, Yi-lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. [Not all errors are equal: Learning text generation metrics using stratified error synthesis](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [Bartscore: Evaluating generated text as text generation](#).
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert](#).
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. [MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.
- Michał Ziemska, Marcin Junczys-Dowmunt, and Bruno Poulliquen. 2016. [The United Nations parallel corpus v1.0](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Appendix

B Quality Prediction Model

We initialized our quality prediction model with pretrained masked language model (e.g. XLM-R). During training, SEScore2 takes in raw text x_0 and synthesized text x_k , supervised with regression score s_k . We drive the sentence embedding from the average pooling of the last layer. Inspired by the prior approach (Shimanaka et al., 2018), we extract two features between sentence embeddings x_0 and x_n : 1) element-wise sentence product and 2) element-wise sentence difference. We concatenated above two features into one vector and feed into a neural network regressor. During inference, when given unseen candidate and reference pair, SEScore2 can directly output a regression score.

C Index Table Construction

We use LASER library⁶ to compute all the sentence embeddings and use Faiss library⁷ to build the index table for English, German and Japanese. We used 8*A100 GPUs for the index table construction. The duration for building index table for English, German and Japanese is 48 hours, 24 hours and 48 hours respectively. From the constructed index table, we extracted 128 nearest neighbors for each raw text. To ensure our learned metrics can cover diverse domains and tasks, we sample millions of raw sentences from diverse domains of corpuses and build ten-million scale index tables. Detailed statistics are discussed at Section 5.1.1.

D Margin-based Criterion

We follow the implementation of margin criterion in (Schwenk et al., 2021). We set the threshold of margin criterion to be 1.06 and extract 128 nearest neighbors to estimate mutual translation capability.

E Pretrained Model Initialization

Since our pipeline utilized pretrained model, we try to answer the question that with the setting, can different pretrained model initialization lead to different performance? In particular, we studied two prior used pretrained models: RemBERT (used by BLEURT) and XLM-R (used by COMET). Based on prior study (Chung et al., 2020a), RemBERT empirically outperforms XLM-R over multiple multi-

⁶<https://github.com/facebookresearch/LASER>

⁷<https://github.com/facebookresearch/faiss>

	Zh→En	WMT21News	WMT22News	Chat	E-comm	Social	TED	Overall
With	BLEURT	0.357	0.350	0.301	0.365	0.371	0.224	0.330
	COMET(DA)	0.360	0.317	0.301	0.330	0.335	0.220	0.312
Without Supervision	BLEU	0.176	0.049	0.117	0.159	0.171	0.092	0.130
	ChrF	0.201	0.045	0.151	0.174	0.198	0.124	0.151
	TER	0.210	-0.049	-0.110	-0.109	-0.158	0.136	0.003
	YiSi	0.302	0.246	0.242	0.315	0.322	0.195	0.273
	BERTScore	0.296	0.268	0.264	0.326	0.333	0.199	0.282
	PRISM	0.285	0.197	0.214	0.259	0.257	0.194	0.238
	BARTScore	0.262	0.149	0.184	0.175	0.195	0.154	0.190
	SEScore	0.334	0.245	0.303	0.352	0.337	0.228	0.301
	SEScore2	0.340	0.276	0.310	0.359	0.328	0.257	0.313

Table 6: Segment-level Kendall correlation on Zh→En at WMT21 News, TED, WMT22 News, Chat, E-commerce, Social, TED and overall performance. SESCOre2 improves over SESCOre by 4% in overall Kendall correlation.

	En→De	WMT21News	WMT22News	Chat	E-comm	Social	TED	Overall
With	BLEURT	0.252	0.387	0.345	0.354	0.333	0.252	0.308
	COMET(DA)	0.239	0.378	0.316	0.316	0.304	0.259	0.292
Without Supervision	BLEU	0.083	0.174	0.210	0.183	0.134	0.113	0.142
	ChrF	0.114	0.210	0.264	0.218	0.176	0.146	0.180
	TER	0.098	-0.119	-0.222	-0.175	-0.159	0.131	-0.045
	YiSi	0.172	0.285	0.225	0.278	0.221	0.212	0.222
	BERTScore	0.169	0.251	0.265	0.251	0.204	0.189	0.215
	PRISM	0.192	0.263	0.195	0.248	0.247	0.238	0.225
	BARTScore	0.017	0.076	0.131	0.138	0.062	0.067	0.075
	SEScore	0.211	0.296	0.237	0.310	0.262	0.241	0.250
	SEScore2	0.227	0.302	0.250	0.268	0.276	0.258	0.260

Table 7: Segment-level Kendall correlation on En→De at WMT21 News, TED, WMT22 News, Chat, E-commerce, Social, TED and overall performance. SESCOre2 improves over SESCOre by 4% in overall Kendall correlation.

Initialization	Zh→En		En→De	
	News	TED	News	TED
SEScore2 (XLM-R)	0.340	0.257	0.206	0.249
SEScore2 (RemBERT)	0.348	0.271	0.227	0.258

Table 8: Segment-level Kendall correlation using different pretrained model initialization on WMT21 En-De and Zh-En at both News and TED domains.

lingual downstream tasks. In Table 8, we demonstrate that compared to XLM-R initialization, our SESCOre2 with RemBERT initialization can further improve Kendall correlations in all language directions. This finding suggests that SESCOre2 with a better pretrained model initialization can increase its learning capacity of score distribution and improve its correlations to human ratings.

F How to interpret SESCOre2?

In order to interpret the evaluation aspects of the SESCOre2, we conducted multi-dimensional human correlations in WebNLG and BAGEL benchmarks. In Table 9, we observe that SESCOre2 achieves the

highest segment-level Kendall correlation across all six quality aspects. SESCOre2’s overall quality score is most correlated to fluency and text structure, which leads the second highest metric BLEURT by 2.5% and 2% Kendall correlation, respectively. In Table 10, we observe that SESCOre2 outperforms all metrics significantly in terms of naturalness and quality. To conclude, despite of producing only a single score, SESCOre2 can be a great indicator of diverse quality aspects of text generation. Specifically, SESCOre2 achieves higher correlation in the quality and fluency aspects of quality estimation.

G Domains in WMT22

In the WMT22 shared metric task (Freitag et al., 2022b), organizers further include three more domains other than News, including chat, E-commerce and social. In Table 6, from Chinese-to-English, SESCOre2 achieves the overall highest Kendall correlation among all the unsupervised metrics. It even outperforms the supervised metric COMET by 0.3% and achieves close performance to the SOTA supervised BLEURT. Comparing to

WebNLG Data-to-Text Generation					
Model Name	Cor	Cov	Flu	Rel	Str
TER	-0.075*	-0.060*	-0.082*	-0.067*	-0.082*
BLEU	0.077*	0.062*	0.075*	0.065*	0.070*
ChrF	0.088*	0.087*	0.082*	0.076*	0.073*
BARTScore	0.096*	0.085*	0.107*	0.079*	0.102*
BERTScore	0.141*	0.110*	0.143*	0.108*	0.142*
SEScore	0.138*	0.114*	0.150*	0.108*	0.139*
PRISM	0.146*	0.121*	0.154*	0.117*	0.143*
BLEURT	0.155	0.128*	0.154*	0.117*	0.148*
SEScore2	0.157	0.144	0.179	0.135	0.168

Table 9: Segment-level Kendall Correlation on WebNLG Data-to-Text generation. * indicates that SESCOre2 significantly outperforms the baseline metric ($p < 0.05$). Cor, Cov, Flu, Rel and Str represents Correctness, Coverage, Fluency, Relevance and Text Structure respectively.

BAGEL Dialogue Generation			
Model Name	Informativeness	Naturalness	Quality
TER	-0.055*	-0.127*	-0.079*
chrF	0.182*	0.078*	0.064*
BLEU	0.138*	0.104*	0.085*
BERTScore	0.217*	0.114*	0.159*
BARTScore	0.183*	0.114*	0.183*
SEScore	0.205*	0.187*	0.184*
PRISM	0.225	0.184*	0.184*
BLEURT	0.254	0.188*	0.180*
SEScore2	0.225	0.204	0.199

Table 10: Segment-level Kendall Correlation on BAGEL dialogue generation. * indicates that SESCOre2 significantly outperforms the baseline metric ($p < 0.05$).

COMET, the supervised learned metrics which only trained on News domain data, SESCOre2 outperforms COMET in three out of four non-news domains, with average 9.6% Kendall correlation improvements. In contrast to SESCOre, another strong unsupervised learned metric baseline, SESCOre2 outperforms SESCOre in 5 out of 6 domains, with overall 3.3% Kendall correlation improvements. In Table 7. for English-to-German, SESCOre2 achieves the highest Kendall correlation among all unsupervised metrics. SESCOre2 achieves the close performance to the supervised learned metrics. SESCOre2 outperforms SESCOre in four out of six domains, with 4% overall Kendall correlation improvements. This result demonstrates that the robustness and generalization of SESCOre2 across domains in both language directions. To understand the failure cases of SESCOre2, we conduct detailed analysis of SESCOre2 at social domain of Zh→En, chat domain of En→De and E-commerce domain of En→De in Section I.

H Retrieval Augmented Examples

I Error Analysis of SESCOre2