MuLER: Detailed and Scalable Reference-based Evaluation

Taelin Karidi
Hebrew University of Jerusalem
taelin.karidi@mail.huji.ac.il

Leshem Choshen
Hebrew University of Jerusalem
leshem.choshen@mail.huji.ac.il

Gal Patel
Hebrew University of Jerusalem
gal.patel@mail.huji.ac.il

Omri Abend
Hebrew University of Jerusalem
omri.abend@cs.huji.ac.il

Abstract

We propose a novel methodology (namely, **MuLER**) that transforms any reference-based evaluation metric for text generation, such as machine translation (MT) into a fine-grained analysis tool. Given a system and a metric, MuLER quantifies how much the chosen metric penalizes specific error types (e.g., errors in translating names of locations). MuLER thus enables a detailed error analysis which can lead to targeted improvement efforts for specific phenomena. We perform experiments in both synthetic and naturalistic settings to support MuLER's validity and showcase its usability in MT evaluation, and other tasks, such as summarization. Analyzing all submissions to WMT in 2014−2020, we find consistent trends. For example, nouns and verbs are among the most frequent POS tags. However, they are among the hardest to translate. Performance on most POS tags improves with overall system performance, but a few are not thus correlated (their identity changes from language to language). Preliminary experiments with summarization reveal similar trends.[1]

1 Introduction

Reference-based evaluation of text generation plays a uniquely important role in the development of machine translation (Papineni et al., 2002), summarization (Lin, 2004), and simplification (Xu et al., 2016) among many other sub-fields of NLP. It allows a scalable, cheap evaluation that often correlates at the system-level with human evaluation.

However, reference-based evaluation metrics tend to produce a bottom line score, allowing little to no ability for a fine-grained analysis of the systems' strengths and weaknesses. Such an analysis is important, for example, for targeted development efforts that focus on improving specific phenomena, or for better identifying scenarios in which the system is reliable (Liu et al., 2021). We propose a novel evaluation methodology, **Multi-Level Evaluation with Reference** (MuLER), that presents a detailed picture of text generation system's performance. Our methodology allows to slice the data according to different criteria, such as syntactic or semantic ones. Given a feature that can be detected automatically on the target side, and a reference-based metric, MuLER allows to scalably measure the system's performance on words and spans that contain this feature.

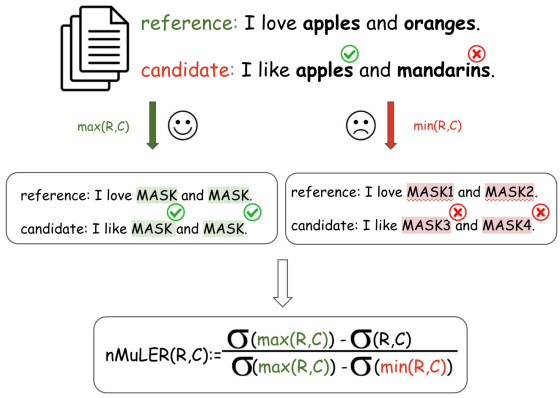


Figure 1: Illustration of MuLER for the feature NOUN. Two masking strategies are employed on the reference and the candidate – Oracle masking $max(R, C)$, and anti-oracle masking $min(R, C)$. σ is the task's metric (e.g. BLEU, ROUGE).

MuLER thus yields a decomposition of any evaluation metric, to more focused measurements of the system's performance on span-level and word-level features, such as POS tags, named entity types, sentence sentiment etc. Moreover, the methodology and code can be expanded to features of choice.

In providing a per-phenomenon picture of system performance, MuLER is similar to challenge set approaches to evaluation (see §6). However, MuLER takes a more naturalistic approach, and narrows the evaluation to the test examples that

[1]Our codebase is found here: https://github.com/tai314159/MuLER

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 436–455
December 6–7, 2023.

contain a particular feature.

Given an evaluation metric (e.g., BLEU) for a text generation task (e.g., MT) and a feature of interest of the system's output (e.g., performance on adjectives), MuLER operates as follows (see §2.1): It masks the feature in both the reference and the prediction by the same token (e.g., replace each adjective with a placeholder "ADJ"). This can be seen as an oracle adaptation to the output, that changes the span with the feature to agree with the reference. MuLER's score is the (normalized) difference between the metric score over the masked texts and the score over the original ones.

We present results from MT as well as summarization and synthetic paraphrasing. In addition, we perform synthetic experiments to validate MuLER's effectiveness and usability. Our experiments show that MuLER can measure performance on a particular feature (§5), and reveal some previously unreported patterns in established MT systems (§4). For example, while translation of nouns and verbs improved over the years, translation of named entities improve only for some categories §4.2.

2 Methodology

The **MuLER** methodology seeks to gain insight as to the performance of a text generation system s according to a given metric σ on instances with the feature f. The feature is a dimension along which the system is evaluated that can be automatically detected given text. Examples of features here may be POS tags, named entity types, morphological categories, among others.

MuLER operationalizes this notion as improvement in the score of s according to σ, if s would have correctly predicted all instances of this feature. For scale, this improvement is compared to the overall possible improvement (the score is defined in §2.3). To assess that, MuLER creates an oracle where the feature f is perfect and an anti-oracle where it is fully wrong (cf. §2.2).

2.1 Feature Tagger: Formal Definition

Let f be a feature of interest. Let $S = \{s_1, ..., s_n\}$ be a corpus of output sentences (produced by the evaluated system), $R = \{r_1, ..., r_n\}$ be a set of corresponding references, and $C = \{c_1, ..., c_n\}$ be a set of corresponding candidates. Let τ be a function from sentences $x \in S \cup R$ that replaces each span containing a feature f with a special mask token M_f (we assume the spans with the f feature are non-overlapping). Denote the i-th token in $\tau(x)$ with $\tau(x)^{(i)}$. Then, for each token $\tau(x)^{(i)}$:

$$\tau(x)^{(i)} = \begin{cases} M_f & \text{if } x^{(i)} \text{ is part of a span} \\ & \text{with the feature } f \\ x^{(i)} & \text{otherwise} \end{cases} \quad (1)$$

2.2 Oracle and Anti-Oracle Masking

Let σ be a reference-based evaluation metric that takes sets of system outputs S and references and R and returns a real value. We can define two masking strategies that represent the best possible performance on sub-spans marked by f, or the worst performance, by applying τ to S and R.

We refer to the optimistic masking strategy as **oracle** masking and denote it by

$$\tau_{\max}(s_1, s_2) = (\tau(s_1), \tau(s_2)).$$

This strategy coincides with eq. 1. For example, if we take f to be common nouns:

Reference: John likes apples and oranges.
Output: John loves bananas and apples.

$\tau_{\max}$(reference) = John likes **NOUN** and **NOUN**.
$\tau_{\max}$(output) = John loves **NOUN** and **NOUN**.

To minimize rather than maximize $\sigma(R, C)$ by masking spans with the feature f, we apply different masks to the outputs and the references. This strategy generally decreases σ, as it deletes existing correspondences between the reference and the outputs. We refer to this masking strategy as **anti-oracle** masking and denote it with $\tau_{\min}$.

Repeating the example above (NOUN and NOUN' are different tokens):

reference: John likes apples and oranges.
output: John loves bananas and apples.

$\tau_{\min}$(reference) = John likes **NOUN** and **NOUN**.
$\tau_{\min}$(output) = John loves **NOUN'** and **NOUN'**.

Let $I \subseteq \{1, ..., n\}$ be the indices for which both $r_i \in R$ and $c_i \in C$ contain a span with the feature f. The average score with each oracle would be:

$$\max_\sigma(R, C) := \frac{1}{|I|} \sum_{i \in I} \sigma(\tau_{max}(r_i, c_i),$$

$$\min_\sigma(R, C) := \frac{1}{|I|} \sum_{i \in I} \sigma(\tau_{min}(r_i, c_i).$$

2.3 MuLER Score

Using these definitions, we may now define the MuLER score. We define the **MuLER score** as:

$$MuLER(R, C) := \frac{\max_\sigma(R, C) - \sigma(R, C)}{\max_\sigma(R, C) - \min_\sigma(R, C)} \quad (2)$$

We compute MuLER variants only on indices in which both the reference and the output contain f (prevents division by zero). Note that lower MuLER score indicates better performance.

Intuitively, MuLER captures the potential gains obtained by the best f, where the numerator of the score captures the absolute gains from improving f. MuLER is therefore a unitless metric, that measures how much of the potential gain is realized by improving the generated spans with the feature f.

For simplicity of notation, we assume a single reference per sentence, but the formulation generalizes straightforwardly to multi-reference settings.

2.4 Normalization Term: Discussion

In this section we provide the motivation behind the normalization term in our score (eq. 2). MuLER seeks to assess a system's ability per feature exhibited in the text. Ideally, features could be analyzed both in a single system (§4.1) and across systems (§4.2). However, the latter may require special treatment. To illustrate this claim, imagine two MT systems, one nearly perfect and another that produces random outputs. The perfect system has little to gain by masking spans of a feature f.and hence the numerator of MuLER will be around zero. However, this is also the case for the random system, since there is hardly any margin for improvement. Even if some words are correctly predicted, the malformed context means a low sentence score. This hints that the numerator is not comparable between systems with substantially different performance and therefore should be normalized.

In order to better capture the systems' overall performance, we leverage the anti-oracle masking, noting that $\sigma(R, C)$ is in the interval $[\min_\sigma(R, C), \max_\sigma(R, C)]$ (except for edge cases, App. §7). The length of this max-min interval can be interpreted as the quality in which the system manages to translate the contexts of spans bearing the feature f (the farther the oracle and the anti-oracle are apart, the better the system is in translating the contexts). To illustrate this point, consider the two extremes. For a high performing system the distance between $\min_\sigma(R, C)$ and $\max_\sigma(R, C)$ is expected to be substantial. There is a lot to lose from an error. However, a horrible system will have a small distance as the minimum and the maximum will both be around zero.

2.5 Leveraging Sentence Scorers

Often, instead of a tagger, a continuous scoring function is available for f. A scorer operates on tokens or sentences to capture a certain aspect of the text (such as sentiment or concreteness). We propose a way to utilize scorers to analyze the system's generation abilities along various dimensions.

Let $\sigma : S \to \mathbb{R}$ be a scoring function, where $S = \{s_1, ..., s_n\}$ is a set of sentences. For a set of references $R = \{r_1, ..., r_k\}$ and a set of candidates $C = \{c_1, ..., c_k\}$, where c_i is the candidate of r_i we define a score s_σ the following way:

$$s_\sigma(R, C) := \frac{1}{k}\sum_{i=1}^{k}(\sigma(r_i) - \sigma(c_i)).$$

Complementing scores. MuLER is defined only for sentences in which the reference and the candidate contain the feature f. Hence, it checks the quality of generation but not cases of over/under generation. To account for such cases and ensure the system even generates the feature, we define a **discrepancy breakdown**:

$$\eta(f) = [\eta_1(f), \eta_2(f), \eta_3(f)]$$

The discrepancy breakdown consists of 3 numbers; **add** ($\eta_1(f)$), **hit** ($\eta_2(f)$), and **miss** ($\eta_3(f)$) scores. $\eta_1(f)$ is the number of sentences in which the feature f appears in the reference more times than it appears the output, $\eta_2(f)$ is the number of sentences in which the feature f appears in the output more times than in the reference and $\eta_3(f)$ is the number of other sentences with equal amount of times. See §4.6 for usage example of the score.

3 Experimental Setup

Evaluation Metrics. As reference-based metrics, we consider BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019) and ROUGE (Lin, 2004). BLEU was developed to measure machine translation quality, and focuses on precision. ROUGE is made for summarization and focuses on recall. Both are based on overlapping n-grams, while BERTScore, a metric for text generation quality, is based on similarity between contextualised embeddings. For these metrics, the basic unit of evaluation is a sentence, as it compares between

a reference sentence (a human translation) and a candidate sentence (an output of a system).

Features. We experiment with several feature types, each separated into different features: POS tagging, NER and dependency features (see App. §9 – for full description).

Sentence Scorers. As dedicated scorers, we look at sentiment analysis, concreteness, valence, dominance and arousal (cf. App. A.)

Released Library Specifications. Upon acceptance, we will share a library of code. The library allows using the metrics used in this paper as well as easily defining new ones. It reports MuLER variants as well as discrepancy breakdowns (§2.5).

3.1 Datasets

WMT. We use the official submissions and references from WMT 2014-2020 news translation task (Bojar et al., 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020). We use all language pairs in each year with English as a target language.

Gender. We make use of the WinoGender dataset (Rudinger et al., 2018) where each sentence has a variation of male, female and neutral (App. §E.1).

Paraphraes. We use the Minimal Paraphrase pairs corpus by Patel et al. (2022). It contains parallel corpora with two syntactic variation types: active versus passive sentences and adverbial clause versus noun phrases. The changes to the sentences are minimal, specifically, the semantic meaning remains identical. See App E.1 for more details.

4 Experiments with Naturalistic Data

4.1 Single Model Analysis

A key point of MuLER is the ability to compare the performance of various features on a single model. Such an analysis can reveal the system's strengths and weaknesses and potentially lead to a targeted development effort on specific features, or be used for debugging purposes. It enables the user to decide where to invest his efforts and allows for a more scientifically-oriented investigation of the results. Fig. 2 shows a standard MuLER report for two systems.

4.2 Comparison Across Systems

We compare WMT systems through years, architectures and performance patterns.

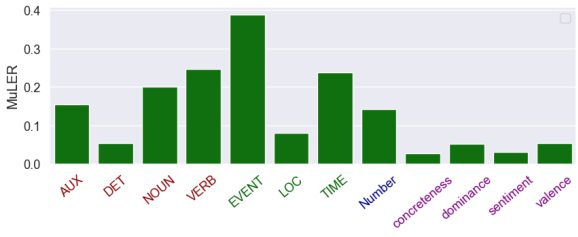


Figure 2: Standard MuLER report. Chinese-English for a subset of features. The Newstest2020 dataset. Submission *Huoshan Translate.919*.

MuLER Similarity to Other Measures. We compute Pearson correlation between negative MuLER scores and BLEU, for every source language, over all submissions of WMT (2014−2017). We use negative MuLER so that high correlation means improvements in both performance measures (e.g., BLEU and MuLER), as reference and candidate similarity is indicated by high BLEU but low MuLER. Fig. 3 shows that BLEU and MuLER are not always correlated. We see that arousal, concreteness, dominance, sentiment and valence scores are in high agreement between MuLER and BLEU. However, some features, e.g., most of the named entity types, are not. This suggests that overall BLEU improvements do not necessarily mean better named entity translations.

We also see that different languages behave differently with respect to the type of features for which MuLER and BLEU are highly correlated. For example, in Chinese, BLEU is more correlated with MuLER, over many different POS tags. This could be explained by differences in the structure of the languages (e.g., syntax). A possible explanation might be that Chinese is simpler to translate in terms of overlapping unigrams (i.e., when syntax is ignored). We do the same analysis comparing MuLER to indices-BLEU (BLEU over the indices in which the feature appears both in the reference and the output) and their $max(R,C) - min(R,C)$ term. We get similar results (see App. 10).

Systems Over Time. We compare WMT systems (see §3.1) from different years and language pairs with MuLER. Overall, there is a consistent trend (see Figs. 4,5,6): as BLEU improves, MuLER improves. However, this trend is not uniform across all features. For certain phenomena, improvement is not consistent with system quality. This is shown by a near-zero or positive correlation between MuLER and the $max(R,C) - min(R,C)$

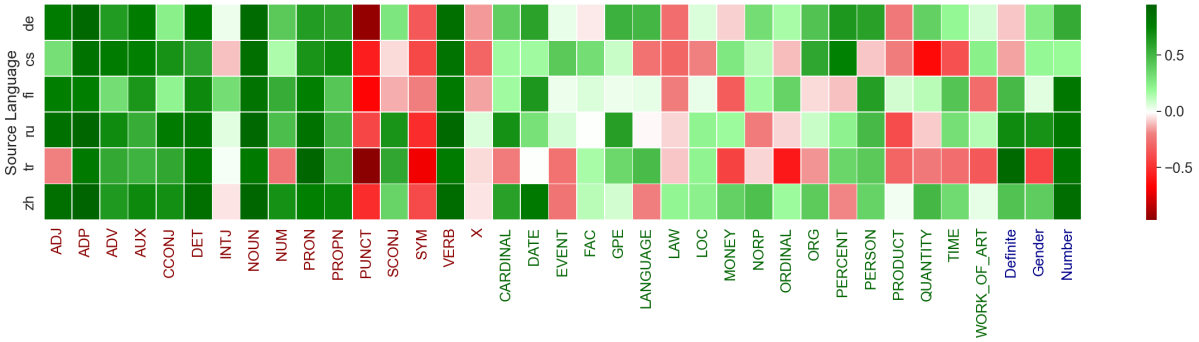


Figure 3: Similarity of Measures. Correlation between BLEU and -MuLER per feature (column) and source language (row). Positive values suggest better systems by BLEU better translate the feature.

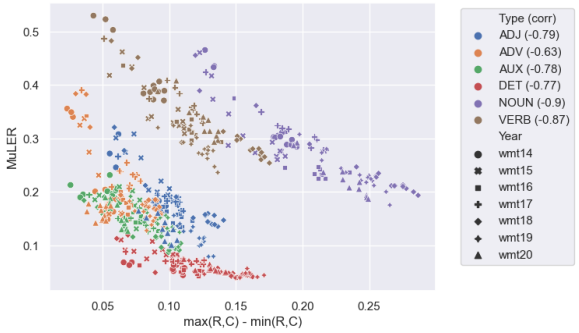


Figure 4: MuLER vs. max(R,C) minus min(R,C) calculated on selected POS-tags. All submissions to WMT (2014 – 2020) for German-English. Next to each POS-tag is the correlation between all x-axis and y-axis values for the POS tag.

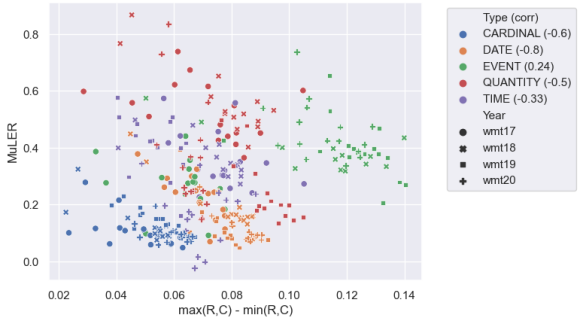


Figure 5: MuLER vs. max minus min calculated on named entities. All submissions to WMT (2017 – 2020) for Chinese-English. Next to each entity is the correlation between x-axis and y-axis values for the entity.

term (indicative of the system's performance on the sentences containing f).

Surprisingly, we find that nouns and verbs are among the hardest POS tags to translate (Fig. 4). On the face of it, this is unexpected, as they account for the most frequent POS tokens in training. Potentially, being open class makes them harder, nouns are common, but each noun by itself is rare. This may also explain why determiners that are frequent are easy and why adverbs are harder than the more frequent auxiliary. Similar trends are presented when comparing MuLER to the total BLEU score of the systems (Fig. 6).

4.3 Manual Analysis

To verify the effectiveness of MuLER, we perform manual analysis and compare pairs of systems that are roughly equal in their overall performance (under BLEU), but greatly differ on a given feature f (under MuLER). We compare 5 pairs of systems and a total of 201 sentences (App. §10).

We consistently see that systems with lower MuLER scores (i.e., better performance) translate feature f better (see Table 1). This means that the neighborhood of f in the candidate sentence is more similar to the reference, not only the masked span itself. Interestingly, we encounter many cases in which the span of f is the same in the reference and both candidates, but the overall translation (i.e., the neighborhood) is better in the one with the lower MuLER. Table 10 shows that out of 97 sentences where quality differs, the system MuLER predicts to be better, indeed translates better in 91.3% of the sentences.

4.4 MuLER with ROUGE: Summarization

We compute MuLER on 3 summarization models (App. §B) and various features. Fig. 7 shows a standard MuLER report, computed under the ROUGE metric. We see that strengths and weaknesses vary between the different systems. Moreover, we see that the concreteness score is always lower than the other scores provided by the sentence scorers (i.e, valence, dominance, arousal and sentiment). Inherently, we expect summarization outputs to be con-

Year	Language pair	Feature type	Feature	Reference	System A	System B
2020	ru-en	POS	AUX	"This **is** heavy oil.	"This **is** thick oil.	"It'**s** thick oil.
2019	fi-en	NER	LOC	Daytime temperatures are between 7 and 12 degrees Celsius, but cooler in **Northern Lapland**.	Daytime temperatures are between + 7 and + 12 degrees, it's cooler in **northern Lapland**.	Daytime temperatures are between + 7 and + 12 degrees, the North is cooler **Lapland**.
2018	tr-en	POS	ORDINAL	**Thirdly**, technology is developing very fast.	**Thirdly**, technology is evolving rapidly.	**Third**, technology is evolving rapidly.
2018	tr-en	POS	ADJ	**Single** digit inflation	Inflation is **single** digits	Inflation is the **only** household
2018	tr-en	POS	ADJ	Clearly, the murders have a **chilling** effect.	The killings clearly had a **chilling** effect.	The killings have clearly had a **cold** shower effect.

Table 1: Example sentences from WMT's submissions. System A has a lower MuLER score than system B. We indicate whether the chosen feature is **consistent** or **inconsistent** with the reference.

synthetic features					features		
average proportion (reference)	average proportion (output)	average MuLER	variance MuLER	std MuLER	feature	average proportion	MuLER
0.22	0.22	0.44	4.09e-04	0.01	NOUN	0.22	0.26
0.15	0.15	0.22	2.24e-04	0.01	VERB	0.12	0.29
0.11	0.11	0.21	6.04e-04	0.03	PROPN	0.09	0.07
0.07	0.07	0.21	2.53e-04	0.02	PRON	0.07	0.16

Table 2: Specificity of MuLER. Comparison of MuLER for synthetic features ("average MuLER") with real features ("MuLER"). The two leftmost columns are the average proportion of the synthetic features in the reference and output. The "average proportion" column indicates the average frequency of the features (e.g, NOUN/VERB) in the reference and the output (as described in §5). WMT 2019 submission; "online-G.0" for German-English.

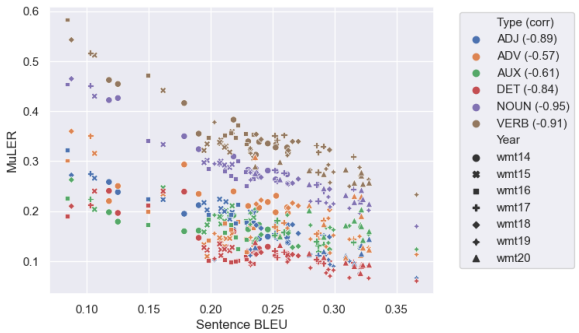


Figure 6: POS-tag MuLER vs. BLEU. All submissions to WMT (2014 – 2020) for Russian-English. Next to each POS-tag is the correlation between all x-axis and y-axis values for the POS-tag.

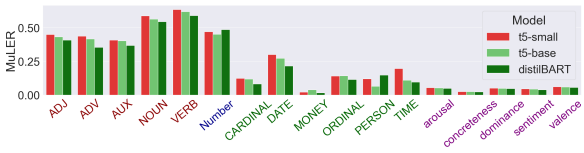


Figure 7: MuLER for summarization. MuLER score is calculated for various features, under ROUGE. We compare 3 models; t5 small, t5 base and distill BART.

crete, as compressing the text is often achieved by simplification. This is indeed revealed by MuLER.

4.5 MuLER with LM-based Metrics

To validate that MuLER could be easily adapted to LM-based metrics, in addition to BLEU, we perform our analysis for the task of MT, also with BERTScore ((Zhang* et al., 2020)). We randomly choose 5 systems from WMT-2020 for Chinese-English. Preliminary experiments show that MuLER can be straightforwardly extended (App. §C) to such metrics.

(a) Clause / Noun Phrase (b) Active / Passive

Figure 8: Discrepancy breakdown of verbs, nouns and auxiliaries for minimal syntactic paraphrases.

4.6 Paraphrases and Gender

We apply MuLER to special cases to demonstrate its usefulness.

Minimal Paraphrases. We compare Minimal Paraphrases (§3.1, App. §E.1) as if they were an output and reference. Evidently, the discrepancy breakdown identifies phrasing differences (see Fig. 8). Adverbial clause sentences have more verbs, while noun phrases have more nouns and thus their miss and hit scores complement each other. The scores also recognize voice changes from active to passive; these require additional auxiliaries while keeping the same verbs and nouns.

WinoGender. Gender choice is critical for many applications. We compare sentences which differ only by gender (§3.1, App. §E.1) as if they were an output and reference. Where sentences with different gender receive a high BLEU score (0.8), the gender feature of MuLER is 1.0 – representing the perfect inability of the systems to translate the correct gender. This shows the strength of MuLER over bottom-line metrics (e.g, BLEU) as it reveals the performance on a specific dimension (gender).

5 Validation Experiments

In this section, we perform various synthetic experiments to check the validity of MuLER. For a given feature f, let $\mathcal{F}$ be the set of words tagged as f (e.g., nouns) under τ, and $\alpha \in [0, 1]$.

Range and Monotonicity of MuLER. We expect MuLER to fall in the interval $[\sigma(min(R, C)), \sigma(max(R, C))]$ and to improve as the quality of translation on the feature f improves (monotonicity). That is, if a system outputs the right translation for α cases of $\mathcal{F}$ (and wrong on $1 - \alpha$ cases accordingly), then we expect $MuLER(R, C) \approx \alpha(\sigma(max(R, C)) - \sigma(min(R, C)))$.

We support this claim using synthetic data experiments. We define a hybrid version of MuLER using a combination of oracle (O) and anti-oracle (AO) masking strategies (§2.1). We split $\mathcal{F}$ into two sets roughly containing α and $1 - \alpha$ of its elements, by partitioning according to sorted first letter. That is, we choose η to be the first letter in the English Alphabet for which the set of all words in $\mathcal{F}$ that start with a-η is of size $\geq \alpha\mathcal{F}$. We split $\mathcal{F}$ to 2 sets; one containing all words that start with the letter a-η, and its complement. We mask α of the occurrences of f using AO-strategy, and the rest using O-strategy, both in the reference and the candidate. This construction emulates a range of systems that improve on f as a function of α.

Tables 4,12, 13 show that this hybrid score is indeed always located according to X in the interval $[\min_\sigma(R, C), \max_\sigma(R, C)]$ (e.g., if $X = 2$ then it's in the middle of the interval).

Specificity of MuLER. We set to verify that MuLER is not sensitive to random features in the text. We expect that features that appear in random subsets of the text with the same frequency will have roughly the same score. To verify this, we create synthetic features with the same frequency in $\mathcal{F}$ as real ones (e.g, nouns/verbs) and compute MuLER over them. Let U be the unique list of words in the union of R and C. For $1 \leq j \leq 1000$: we split U to p equally sized groups $\{U_1, ..., U_p\}$ (we ignore the remainder). Indeed, as seen in Table 2, the average proportion of U_i in R and C is roughly the same. For $1 \leq i \leq p$ we compute $MuLER(R, C)$ by masking only the words in U_i (both in R and C). At each run we have p scores $\{(m_1, ..., m_p)_j\}_{j=1}^{1000}$ from which we choose one randomly. In total, we get 1000 scores: $M = \{\tilde{m}_1, ..., \tilde{m}_{1000}\}$. We compute the variance and standard deviation for M (see Table 2). We find that the variance and std are around zero across values of p, for $p \in \{2, ..., 6\}$ (see App. §14). Meaning, MuLER is not specified to random phenomena. Moreover, the results are different compared to *real* linguistic phenomena with the same frequency (e.g, nouns/verbs, see Table 2). These findings suggest that MuLER is not sensitive to variation that does not reflect variation in quality.

Robustness to Feature Frequency. We start by validating that MuLER score is less sensitive to the frequency of f.

We split $\mathcal{F}$ into two sets roughly containing α

system	50% abl-MuLER		100% abl-MuLER		50% MuLER		100% MuLER	
	noun	verb	noun	verb	noun	verb	noun	verb
Facebook_FAIR.6750	0.021	0.018	0.054	0.034	0.203	0.320	0.267	0.391
online-A	0.023	0.017	0.055	0.036	0.229	0.357	0.295	0.432
UCAM.6461.	0.023	0.017	0.054	0.035	0.220	0.328	0.279	0.405

Table 3: Robustness to Feature Frequency. Presented here are 3 submissions from WMT 2019, translation from German to English (see Table 15 for more results). We compare between MuLER and abl-MuLER (MuLER's numerator – an ablated version of MuLER) with $50\%/100\%$ of nouns/verbs masked.

year	langs	submission	system bleu	bleu indices		MuLER		O		AO		hybrid	
				n	v	n	v	n	v	n	v	n	v
20	de-en	newstest2020.de-en.OPPO.1360	0.39	0.41	0.41	0.18	0.29	0.45	0.45	0.21	0.32	0.33	0.38
18	ru-en	newstest2018.Alibaba.5720.ru-en	0.30	0.30	0.30	0.24	0.32	0.35	0.34	0.14	0.21	0.24	0.27
15	fi-en	newstest2015.uedin-syntax.4006.fi-en	0.12	0.12	0.13	0.38	0.39	0.17	0.16	0.05	0.08	0.10	0.12

Table 4: Range and Monotonicity of MuLER. MuLER scores on nouns ("n") and verbs ("v") in 5 randomly chosen systems from WMT. Oracle ("O") and Anti-Oracle ("AO") masking strategies vs. hybrid masking strategy (described in §5) at $50-50$ split (50% of noun/verb is masked with O-strategy, and the rest with AO-strategy).

and $1-\alpha$ of its elements, by partitioning according to sorted first letter (as explained before). We then mask α of $\mathcal{F}$ and ignore the rest of the instances. This allows us to test MuLER on a feature with similar performance (a random sample of the original feature) but different frequency, namely α frequency of the feature f across $\mathcal{F}$ (this is not true when doing the split at the sentence-level). We see in Table 3 that MuLER is robust to changes in frequencies (of nouns and verbs), compared to abl-MuLER – an ablated version of MuLER which is defined as MuLER's numerator. This holds across various frequencies and features (see Table 15). This suggests that MuLER is a more suitable score for measuring system performance and that its signal is not due to the frequency of the feature (it may play a role, but not a central one).

6 Related Work

Automatic metrics are useful to assess systems and we base our work on them (see §3). Other lines of work study a specific property and propose evaluation measures for it. For example, addressing hallucinations (Kryscinski et al., 2020) or measuring grammaticality (Vadlapudi and Katragadda, 2010). We share the aspiration to a more fine-grained form of evaluation with these works.

There are methods for analyzing performance in a more fine-grained manner. For example, evaluation with minimal changes to the input (Warstadt et al., 2020) and challenge sets (Macketanz et al., 2018; Emelin and Sennrich, 2021). Few methods highlight patterns rather than predefined properties, by contrasting texts (e.g. reference and output) (Gralinski et al., 2019; Lertvittayakumjorn et al., 2021). In a sense, MuLER stands in the middle between those, it highlights a closed set of traits, but it is extendable.

7 Conclusion

We presented a novel methodology (MuLER) to decompose any reference-based score into its fine-grained components, making it possible to obtain a detailed picture of text generation systems' performance, instead of a bottom-line score. MuLER filters and dissects naturalistic data to highlight phenomena in the generated text. We validated MuLER using a set of synthetic experiments (§5). Applying MuLER to off-the-shelf systems shows (§4) that different systems' strengths and weaknesses are varied, even when their overall performance is alike, and detect interesting trends over

the years. Our work creates an avenue for further research into more fine-grained evaluation metrics and provides a tool to understand system behaviour. In future work, we plan to extend MuLER to more complex features such as long-distance syntactic dependencies.

Limitations

Among MuLER appealing traits is its reliance on existing, accepted and easily changed components. It also counts as its limitation, where the base metric is invariant to a trait, MuLER would also be, where masking tagging or scoring is not available (e.g. in endangered languages) the features would not be possible to extract. In general, detecting a feature (e.g. POS tag) is usually harder than evaluating the quality of its generation, MuLER makes this evaluation more accessible.

By definition, MuLER is as good as the tagger that is used to detect a feature of choice. While there is a potential for noise in the process, the taggers used in this paper are known to work well and are indeed vastly used.

We showcase MuLER on BLEU and ROUGE as they are still among the most widely adopted metrics in their respective tasks. The concept of MuLER can be straightforwardly extended to LM-based metrics and we intend to explore it in future work. For now, we shared initial results on BERTScore suggesting this is indeed the case.

For some validations, we use synthetic experiments, that make a well-controlled experiment, but sometimes lack some characteristics of natural data. Overall, we try to evaluate intrinsically, extrinsically by use cases, manually and synthetically to present a full view where the whole is greater than the sum of its parts.

Although we use MuLER to compare between models, it is not clear whether such a comparison is interesting for systems with overall very different performance; if one system's overall performance is very low, then even if it somehow translates a specific feature well, the quality of its output is bad. However, comparing systems with overall similar performance is the more common use case and hence useful; for example, when choosing between systems with top performance to perform a task or for analyzing the differences between systems.

Acknowledgements

This work was supported in part by the Azrieli Fellowship, the Vatat scholarship, the Israel Science Foundation (grant no. 2424/21), and by the Applied Research in Academia Program of the Israel Innovation Authority.

References

Loïc Barrault, Magdalena Biesialska, Ondrej Bojar, Marta Ruiz Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi kiu Lo, Nikola Ljubesic, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt20). In *WMT*.

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.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference*

on Machine Translation: Volume 2, Shared Task Papers, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

F. Gralinski, Anna Wróblewska, Tomasz Stanislawek, Kamil Grabowski, and T. Górecki. 2019. Geval: Tool for debugging nlp datasets and models. In *ACL 2019*.

Benedict C Jones, Lisa M DeBruine, Jessica K Flake, Marco Tullio Liuzza, Jan Antfolk, Nwadiogo C Arinze, Izuchukwu LG Ndukaihe, Nicholas G Bloxsom, Savannah C Lewis, Francesco Foroni, et al. 2021. To which world regions does the valence–dominance model of social perception apply? *Nature human behaviour*, 5(1):159–169.

Christopher S. G. Khoo and Sathik Basha Johnkhan. 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44:491 – 511.

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.

Piyawat Lertvittayakumjorn, Leshem Choshen, Eyal Shnarch, and Francesca Toni. 2021. Grasp: A library for extracting and exploring human-interpretable textual patterns. *arXiv preprint arXiv:2104.03958*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

Charles Egerton Osgood, George J Suci, and Percy H Tannenbaum. 1957. *The measurement of meaning*. 47. University of Illinois press.

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.

Gal Patel, Leshem Choshen, and Omri Abend. 2022. On neurons invariant to sentence structural changes in neural machine translation. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 194–212, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana. Association for Computational Linguistics.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Ravikiran Vadlapudi and Rahul Katragadda. 2010. On automated evaluation of readability of summaries: Capturing grammaticality, focus, structure and coherence. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 7–12, Los Angeles, CA. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Scorers used

In this section, we elaborate on the scorers' use and their origin.

Sentiment. Sentiment Analysis is the process of determining whether a piece of text is positive, negative or neutral. We follow the method of Khoo and Johnkhan (2018) that relies on per word score and a rule-based combination (mainly dealing with negation). The method was shown to outperform other lexicons and to work well without the need for neural networks. We selected this method as it strikes a good balance between accuracy and running time. We defer the application of neural metrics to future work.

We consider 4 token-level scores which we aggregate into a sentence score by averaging. We ignore words that do not appear in the lexicons.

Concreteness. The Concreteness rating of a word represents to which extent a word is concrete, how perceptible is it. For example, a fruit is less concrete than a banana and tomorrow is more concrete than sometime. The lexicon (Brysbaert et al., 2014) contains 40K lemmas each with a concreteness score.

Valence Arousal and Dominance. In psychology, it is common to discuss three characteristics in how we perceive others (e.g., in recognizing faces (Jones et al., 2021)): valence (pleasure vs. displeasure), arousal (active vs. passive), and dominance (dominant vs. submissive). These were shown to be mostly independent directions of word meaning (Osgood et al., 1957; Russell, 1980, 2003). The lexicon (Mohammad, 2018) contains 20K words and their respective scores for each of those axes.

B Summarization

We compare T5-base (Raffel et al., 2020), T5-small and distillbart (Shleifer and Rush, 2020; Lewis et al., 2020) models on the CNN Daily Mail summarization dataset (Nallapati et al., 2016).

We use models from the HuggingFace model hub. DistillBart-"sshleifer/distilbart-cnn-12-6" and T5-"t5-base" and "T5-small"

C LM-based Metrics

We perform preliminary experiments using BERTScore, which is a language-model (LM) based metric for measuring the quality of generation tasks. We use it together with "bert-based-uncased" model. In order to adapt BERTScore to MuLER, we perform alterations to the similarity matrix of the reference and candidate embeddings, that is calculated during the score's

computation. To compute $\max_\sigma(R, C)$, after the similarity matrix between the un-masked reference and un-masked candidate is computed, we set the ij-th entry to be 1 if both the i-th word in the reference and the j-th word in the candidate is masked (if the masked word is split to multiple tokens by the BERT tokenizer, we set the corresponding entry in the similarity matrix to be 1 for each of them). To compute $\min_\sigma(R, C)$, after the similarity matrix between the un-masked reference and un-masked candidate is computed, we set the i-th row to be zeroes if the i-th word in the reference is masked, and the j-th column to be zeroes if the j-th word in the candidate is masked. Indeed, in this setting we also get that $\min_\sigma(R, C) > \min_\sigma(R, C)$ (this is true for 1000 randomly sampled sentences from the submissions we analyzed). We randomly sampled 5 submissions to WMT-2020 for Chinese-English (Tencent_Translation.1249, Online-B.1605, DeepMind.381, Huoshan_Translate.919 and OPPO.1422). Similar trends to the results obtained by MuLER with BLEU are exhibited.

D Data

We provide the complete MuLER database containing the results for WMT submissions ($2014-2020$) on all features (see 3) in the supplementary materials (App. §E). We will release it together with our code upon acceptance.

E Supplementary Materials

The complete MuLER database (scores for all WMT's submissions ($2014-2020$)) and the tagged manual analysis are in the supplementary materials submitted with the paper.

E.1 Minimal Paraphrases

Minimal Paraphrases dataset (Patel et al., 2022) contains 1169 active-passive pairs and 114 clause-noun phrase pairs. Examples are in table 5.

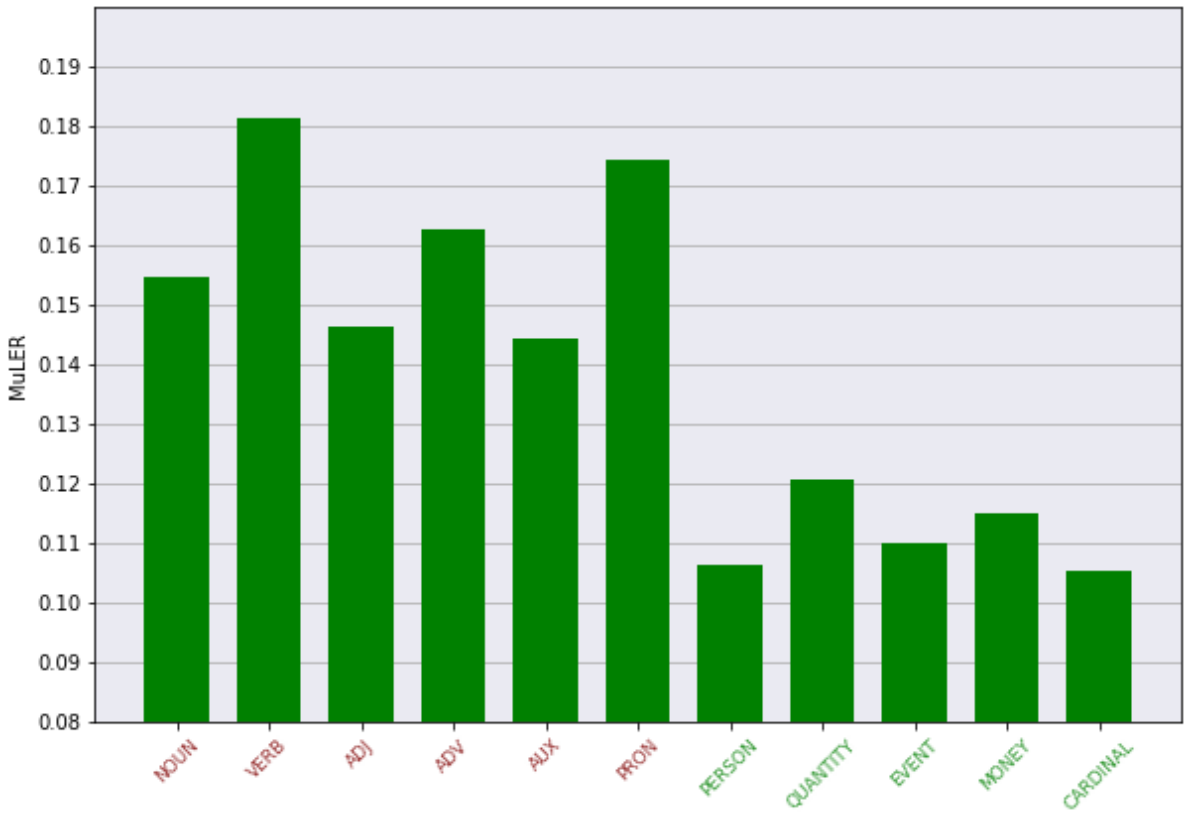


Figure 9: Standard MuLER Report with BERTScore. Chinese-English for a subset of features. The Newstest2020 dataset. Submission "Huoshan Translate.919". MuLER computed with BERTScore.

	Source	Paraphrased
Active Voice→ Passive Voice	*She **took** the book*	*The book **was taken** by her*
Adverbial Clause→ Noun Phrase	*The party died down before **she arrived***	*The party died down before **her arrival***

Table 5: Examples of minimal paraphrases

*The technician told the customer that **she** could pay with cash.* *The technician told the customer that **he** could pay with cash.*
*The supervisor gave the employee feedback on **her** stellar performance.* *The supervisor gave the employee feedback on **his** stellar performance.*
*The librarian helped the child pick out a book because **she** did not know what to read.* *The librarian helped the child pick out a book because **he** did not know what to read.*

Table 6: Female-Male pairs from the WinoGender dataset

E.2 WinoGender

WinoGender (Rudinger et al., 2018) cosists of sentences that differ only by the gender of one pronoun in the sentence, see examples in Table 6.

F Manual Analysis

We perform a small-scale manual analysis to validate MuLER does indicate the quality of performance on a certain feature. We chose 5 systems from different years and language pairs (see Table 10 for full details). We compare pairs of systems that are roughly equal in their overall performance (under BLEU), but greatly differ on a given feature f, under MuLER (see §4.3). One of the authors annotated the data. For every pair of submissions, the data was shuffled such that the sentences were side by side without knowing in advance which is the better system.

G Negative MuLER

Intuitively, we expect to always gain by masking a certain proportion of a given feature in the text (i.e, positive MuLER score). However, there are edge cases in which $max(R, C) - BLEU(R, C)$ is negative. It can be due to a mistake of the tagger or the sentence structure (for example, a word in the reference that is a noun is used in the candidate as a verb, etc.). In table 7 we present examples for such cases.

reference	masked reference	output	masked output
Nitromethane is being used for **example** in **drag racing**.	**NOUN** is being used for **NOUN** in **NOUN NOUN**.	**Nitromethane** is used, for **example**, drag **racing**.	**NOUN** is used, for **NOUN**, drag **NOUN**.
The **film** will premiere in Finland in September 2015.	The **NOUN** will premiere in Finland in September 2015.	The **film** will have its Finnish **premiere** in September 2015.	The **NOUN** will have its Finnish **NOUN** in September 2015.
Its unpredictability unsettled **people**'s **nerves**.	Its unpredictability unsettled **NOUN**'s **NOUN**.	Its **unpredictability** made **people** nervous.	Its **NOUN** made **NOUN** nervous.
Our whole **house** moved, we were trembling with **fear**.	Our whole **NOUN** moved, we were trembling with **NOUN**.	We need the **whole** of our **house** moved: **vapisimme fear**.	We need the **NOUN** of our **NOUN** moved: **NOUN NOUN**.

Table 7: Negative MuLER.

H graphs

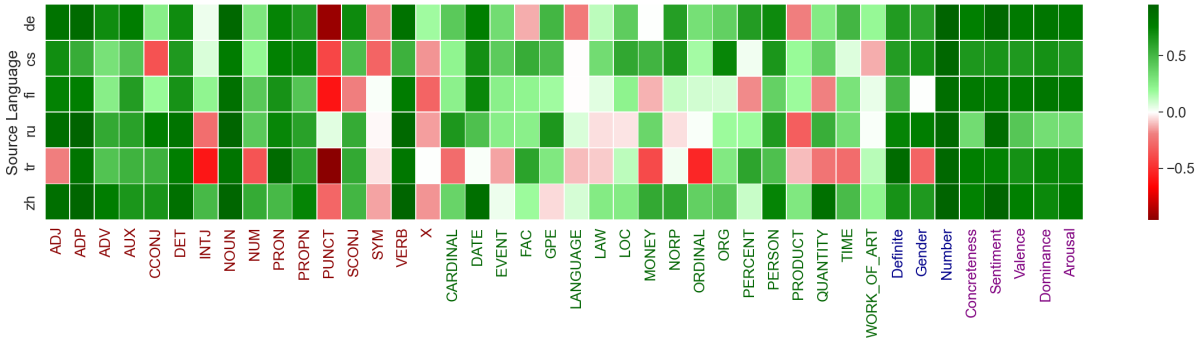
We supply here multiple graphs that were mentioned in the text. The rest of the analysis graphs could be found in the supplementary files.

Year	Languages	Feature type	Feature	Reference	System A	System B
2020	ru-en	POS	AUX	"This **is** heavy oil.	"This **is** thick oil.	"It'**s** thick oil.
2018	tr-en	POS	ORDINAL	**Thirdly**, technology is developing very fast.	**Thirdly**, technology is evolving rapidly.	**Third**, technology is evolving rapidly.
2018	tr-en	POS	ORDINAL	The **first** part was the repairing of the mosque, the main building.	The **first** part was the repair of the mosque, the main building.	The **first** part was the renovation of the main building.
2018	tr-en	POS	ADJ	**Single** digit inflation	Inflation is **single** digits	Inflation is the **only** household
2018	tr-en	POS	ADJ	Clearly, the murders have a **chilling** effect.	The killings clearly had a **chilling** effect.	The killings have clearly had a **cold** shower effect.
2019	fi-en	NER	LOC	Daytime temperatures are between 7 and 12 degrees Celsius, but cooler in **Northern Lapland**.	Daytime temperatures are between + 7 and + 12 degrees, it's cooler in **northern Lapland**.	Daytime temperatures are between + 7 and + 12 degrees, the North is cooler **Lapland**.
2019	fi-en	NER	LOC	It was still peaceful at least in **Crete**, she said early on Saturday evening.	It was still peaceful, at least in **Crete**, "he said on Saturday at the beginning of the evening.	At least there was still calm in **Crete**, "he told the crowd in the early evening on Saturday.

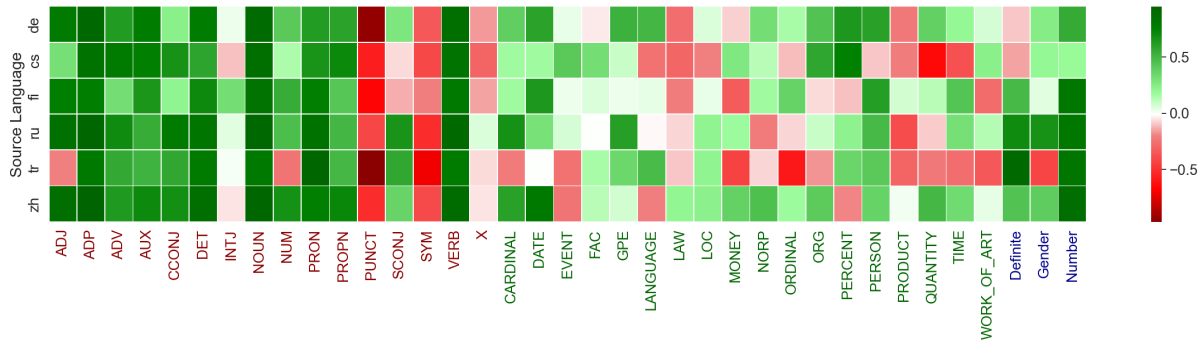
Table 8: Example sentences from WMT's submissions. System A has a lower MuLER score than system B. We indicate whether the chosen feature is **consistent** or **inconsistent** with the reference.

POS tags	named entities	features
NOUN	TIME	GENDER
VERB	WORK_OF_ART	DEFINITE
PUNCT	PERSON	NUMBER
PROPN	NORP	
INTJ	CARDINAL	
NUM	MONEY	
PRON	EVENT	
SYM	ORDINAL	
SCONJ	DATE	
ADJ	FAC	
ADP	ORG	
ADV	LAW	
AUX	PRODUCT	
X	PERCENT	
CCONJ	QUANTITY	
DET	LANGUAGE	
	GPE	
	LOC	

Table 9: Features we use in the paper.



(a) MuLER vs. System Sentence BLEU



(b) MuLER vs. Max BLEU - Min BLEU

Figure 10: Similarity of Measures. Represents correlation of score achievements, e.g. positive values between BLEU and MuLER suggest that BLEU increases as MuLER decreases and vice versa.

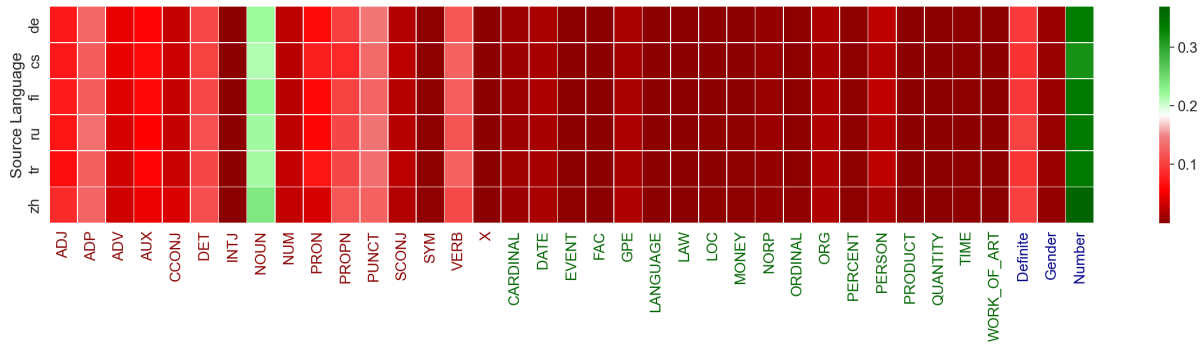


Figure 11: Frequency of MuLER entities. For each language pair we chose the submission with the best BLEU score (from WMT 2014 – 2020) and calculated the average frequency for each feature.

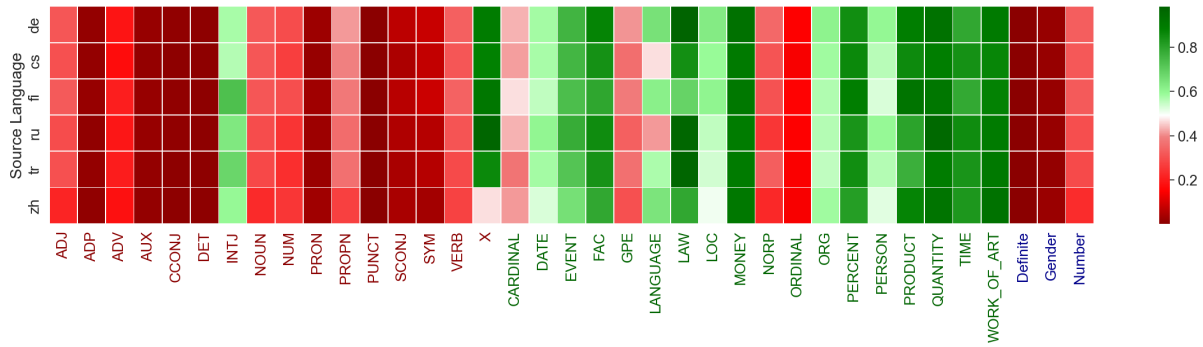


Figure 12: Uniqeness of MuLER entities. For each language pair we choose the submission with the best BLEU score (from WMT 2014 – 2020). For each feature we calculate its average uniqueness, defined as the number of unique times the feature appears in the text, divided by the total times it appears in the text.

year	L1-L2	feature	system A	system B	A=B	A>B	B>A	MuLER A	MuLER B	BLEU indices A	BLEU indices B
19	fi-en	LOC	newstest2019. GTCOM-Primary. 6946.fi-en	newstest2019. USYD. 6995.fi-en	30	23	1	0.30	0.32	0.18	0.32
18	tr-en	ORDINAL	newstest2018. online-A.0.tr-en	newstest2018. online-G.0.tr-en	11	13	0	0.06	0.15	0.22	0.24
20	ru-en	AUX	newstest2020.ru-en.Online-G. 1567	newstest2020.ru-en.eTranslation. 686	31	16	3	0.14	0.20	0.34	0.34
20	zh-en	PERSON	newstest2020.zh-en.OPPO.1422	newstest2020.zh-en.zlabs-nlp.1176	23	37	2	0.17	0.49	0.22	0.19
18	tr-en	WORK_ OF_ ART	newstest2018. online-G.0.tr-en	newstest2018. online-G.0.tr-en	2	6	2	0.01	0.44	0.25	0.26

Table 10: Manual Analysis. system A is the system with a lower MuLER score (i.e, better performance on the feature). *A=B/A>B/A<B* indicates the number of sentences where the translation of the feature was of the same quality between system A and B (or better/worse accordingly). *BLEU indices A/B* is the BLEU score of system A/B on sentences in the reference and the output that contain the feature.

year	langs	submission	system bleu	bleu indices		MuLER		O		AO		hybrid	
				noun	verb	noun	verb	noun	verb	noun	verb	noun	verb
20	de-en	newstest2020.de-en.OPPO.1360	0.39	0.41	0.41	0.18	0.29	0.45	0.45	0.21	0.32	0.33	0.38
15	fi-en	newstest2015.uedin-syntax.4006.fi-en	0.12	0.12	0.13	0.38	0.39	0.17	0.16	0.05	0.08	0.10	0.12
18	ru-en	newstest2018.Alibaba. 5720.ru-en	0.30	0.30	0.30	0.24	0.32	0.35	0.34	0.14	0.21	0.24	0.27
19	de-en	newstest2019.RWTH_ Aachen_System.6818.de-en	0.33	0.33	0.33	0.21	0.28	0.39	0.37	0.14	0.24	0.26	0.30
20	ru-en	newstest2020.ru-en.Online-G.1567	0.32	0.33	0.33	0.22	0.26	0.38	0.36	0.13	0.22	0.26	0.28

Table 11: Range and Monotonicity of MuLER. Presented here are MuLER scores on nouns and verbs in 5 randomly chosen systems from WMT. Oracle (O) and Anti-Oracle (AO) masking strategies vs. hybrid masking strategy (as described in §5) at $50-50$ split (50% of noun/verb is masked with O-strategy, and the rest with AO-strategy).

year	langs	submission	system bleu	bleu indices		MuLER		O		AO		hybrid	
				noun	verb	noun	verb	noun	verb	noun	verb	noun	verb
20	de-en	newstest2020.de-en.OPPO.1360	0.39	0.41	0.41	0.18	0.29	0.45	0.45	0.21	0.32	0.31	0.36
15	fi-en	newstest2015.uedin-syntax.4006.fi-en	0.12	0.12	0.13	0.38	0.39	0.17	0.16	0.05	0.08	0.10	0.12
18	ru-en	newstest2018.Alibaba.5720.ru-en	0.30	0.30	0.30	0.24	0.32	0.35	0.34	0.14	0.21	0.23	0.26
19	de-en	newstest2019.RWTH_Aachen_System.6818.de-en	0.33	0.33	0.33	0.21	0.28	0.39	0.37	0.14	0.24	0.25	0.30
20	ru-en	newstest2020.ru-en.Online-G.1567	0.32	0.33	0.33	0.22	0.26	0.38	0.36	0.13	0.22	0.25	0.28

Table 12: Range and Monotonicity of MuLER. Presented here are MuLER scores on nouns and verbs in 5 randomly chosen systems from WMT. Oracle (O) and Anti-Oracle (AO) masking strategies vs. hybrid masking strategy (as described in §5) at $40-60$ split (40% of noun/verb is masked with O-strategy, and the rest with AO-strategy

year	langs	submission	system bleu	bleu indices		MuLER		O		AO		hybrid	
				noun	verb	noun	verb	noun	verb	noun	verb	noun	verb
20	de-en	newstest2020.de-en.OPPO.1360	0.39	0.41	0.41	0.18	0.29	0.45	0.45	0.21	0.32	0.31	0.36
15	fi-en	newstest2015.uedin-syntax.4006.fi-en	0.12	0.12	0.13	0.38	0.39	0.17	0.16	0.05	0.08	0.09	0.11
18	ru-en	newstest2018.Alibaba.5720.ru-en	0.30	0.30	0.30	0.24	0.32	0.35	0.34	0.14	0.21	0.22	0.26
19	de-en	newstest2019.RWTH_Aachen_System.6818.de-en	0.33	0.33	0.33	0.21	0.28	0.39	0.37	0.14	0.24	0.24	0.29
20	ru-en	newstest2020.ru-en.Online-G.1567	0.32	0.33	0.33	0.22	0.26	0.38	0.36	0.13	0.22	0.24	0.28

Table 13: Range and Monotonicity of MuLER. Presented here are MuLER scores on nouns and verbs in 5 randomly chosen systems from WMT. Oracle (O) and Anti-Oracle (AO) masking strategies vs. hybrid masking strategy (as described in §5) at $30-70$ split (30% of noun/verb is masked with O-strategy, and the rest with AO-strategy

synthetic features							features		
average proportion (reference)	**average proportion (output)**	**variance of average proportion (reference)**	**variance of average proportion (output)**	**average MuLER**	**variance MuLER**	**std MuLER**	**feature**	**average proportion**	**MuLER**
0.22	0.22	4.61e-04	2.57e-04	0.44	4.09e-04	0.01	NOUN	0.22	0.26
0.15	0.15	4.86e-04	7.25e-04	0.22	2.24e-04	0.01	VERB	0.12	0.29
0.11	0.11	3.39e-04	2.90e-04	0.21	6.04e-04	0.03	PROPN	0.09	0.07
0.07	0.07	7.33e-04	7.12e-04	0.21	2.53e-04	0.02	PRON	0.07	0.16
0.05	0.05	6.71e-04	2.07e-04	0.19	6.15e-04	0.02	ADV	0.04	0.18

Table 14: Specificity of MuLER. Comparison of MuLER for synthetic features ("average MuLER") with real features ("MuLER"). The two leftmost columns are the average proportion of the synthetic features in the reference and output. The "average proportion" column indicates the average frequency of the features (e.g, NOUN/VERB) in the reference and the output (as described in §5). Submission is "online-G.0" for German-English from WMT 2019.

system	**50% abl-MuLER**		**100% abl-MuLER**		**50% MuLER**		**100% MuLER**	
	noun	verb	noun	verb	noun	verb	noun	verb
Facebook_FAIR.6750	0.021	0.018	0.054	0.034	0.203	0.320	0.267	0.391
online-A	0.023	0.017	0.055	0.036	0.229	0.357	0.295	0.432
UCAM.6461.	0.023	0.017	0.054	0.035	0.220	0.328	0.279	0.405
uedin.6749	0.022	0.016	0.056	0.034	0.242	0.374	0.306	0.448
online-A	0.023	0.017	0.055	0.036	0.229	0.357	0.295	0.432
online-B	0.018	0.016	0.047	0.032	0.169	0.286	0.225	0.359
uedin.6749	0.022	0.016	0.056	0.034	0.242	0.374	0.306	0.448

Table 15: Robustness to Feature Frequency. Presented here are 3 submissions from WMT 2019, translation from German to English (see Table 15 for more results). We compare between MuLER and abl-MuLER (MuLER's numerator – an ablated version of MuLER) with $50\%/100\%$ of nouns/verbs masked.