Perturbation-based QE: An Explainable, Unsupervised Word-level Quality Estimation Method for Blackbox Machine Translation

Tu Anh Dinh   tu.dinh@kit.edu
Karlsruhe Institute of Technology, Germany

Jan Niehues   jan.niehues@kit.edu
Karlsruhe Institute of Technology, Germany

Abstract

Quality Estimation (QE) is the task of predicting the quality of Machine Translation (MT) system output, without using any gold-standard translation references. State-of-the-art QE models are supervised: they require human-labeled quality of some MT system output on some datasets for training, making them domain-dependent and MT-system-dependent. There has been research on unsupervised QE, which requires glass-box access to the MT systems, or parallel MT data to generate synthetic errors for training QE models. In this paper, we present *Perturbation-based QE* - a word-level Quality Estimation approach that works simply by analyzing MT system output on perturbed input source sentences. Our approach is unsupervised, explainable, and can evaluate any type of blackbox MT systems, including the currently prominent large language models (LLMs) with opaque internal processes. For language directions with no labeled QE data, our approach has similar or better performance than the zero-shot supervised approach on the WMT21 shared task. Our approach is better at detecting gender bias and word-sense-disambiguation errors in translation than supervised QE, indicating its robustness to out-of-domain usage. The performance gap is larger when detecting errors on a nontraditional translation-prompting LLM, indicating that our approach is more generalizable to different MT systems. We give examples demonstrating our approach's explainability power, where it shows which input source words have influence on a certain MT output word.

1 Introduction

Machine Translation (MT), with the aim of translating text from a source language to a target language, has been increasingly adopted in different real-world scenarios, ranging from translations in healthcare areas to translations in the legal domains (Vieira et al., 2021). In many of these applications, errors in translation output could cause serious harm to the users, e.g., translation errors leading to wrong medical diagnoses in healthcare or wrong judgment in court. Therefore, it is important to let the users know how much they can trust a translation, by providing them with some quality assessment of the MT output. This is not always straightforward due to the lack of gold-standard human translations, or the mismatch between evaluation data and real-world usage. As a result, researchers have been looking into Quality Estimation.

Quality Estimation (QE) is the task of predicting the quality of MT system output without access to reference translations. State-of-the-art QE systems are built in a supervised manner,

where they require human-labeled quality assessment on MT output for training (Rei et al., 2022). This approach has 2 drawbacks: the labeled QE data is costly to obtain, and the trained QE models would only know about the types of error that are presented in the training data. Supervised QE models are likely to underperform in unfamiliar settings (Kocyigit et al., 2022), e.g., when evaluating the output of a new MT system on a new dataset from a different domain. Consequently, there has been research into unsupervised QE, where the human-labeled assessment data is no longer required (Fomicheva et al., 2020b; Tuan et al., 2021). These works either require glass-box information of the MT system (e.g., output log probabilities or attention scores), or a large amount of parallel MT data to create synthetic QE data for training. This is problematic for language pairs with low-resourced MT data, or when the MT system is kept blackbox, which is the current trend of some widely-discussed API-only large language models.

In this paper, we propose an **unsupervised** word-level QE approach to evaluate **blackbox** MT systems, termed Perturbation-based QE. Our motivation is inspired by a known problem: when uncertain, MT systems rely on spurious correlations learnt from the training data to generate translation (Emelin et al., 2020; Savoldi et al., 2021). We assume that, when outputting a translation token, if the MT system relies on too many parts of the source sentence, it is likely that the system is exploiting irrelevant correlations, thus the output token is unreliable. Consider the English → German example: *"My friend has a Ph.D. degree, and now she is a professor."* → *"Meine Freundin hat einen Doktortitel, und sie ist jetzt eine Professorin."*. The translation word *"Freundin"* should only depends on *"friend"* and *"she"*, where *"friend"* indicates the meaning and *"she"* indicates the gender form. The output word being influenced by more source words would indicate that the MT system is focusing on the wrong part of the input sentence.

Broadly speaking, in Perturbation-based QE, we perturb words in the source sentences one by one to find out which source words influence a single output word. If an output word is influenced by too many source words, then it is predicted as a bad translation. Due to its simplicity, our approach does not require human-labeled QE data, nor parallel MT data, nor glass-box access to the evaluated MT system. Additionally, our QE approach comes with explainability power: it shows which source words affect each output word in the translation, thus can be used as an indication of the wrong correlations that is inherent in the MT system.

To summarize, our contributions are as follows:

- Proposing Perturbation-based QE[1]: a simple word-level Quality Estimation approach that is explainable, unsupervised and works with any type of blackbox MT systems, including the API-only large language models (i.e., MT-system-agnostic).

- Experiments showing the advantages of Perturbation-based QE: (1) it has similar or better performance than zero-shot QE, without making use of labeled data of auxiliary language pairs and (2) it is domain-independent and MT-system-independent compared to supervised QE methods: it can better capture out-of-domain gender errors and word-sense-disambiguation errors, especially from an unseen, nontraditional translation-prompting large language model and (3) it is not sensitive to hyperparameters.

- Analysis showing an example use of the explainability power of Perturbation-based QE.

2   Related work

Quality Estimation (QE) aims to predict the quality of Machine Translation (MT) output, either at the sentence level or word level. For word-level QE, the goal is to predict whether each word in the translation is correct. State-of-the-art word-level QE methods are supervised (Kim et al., 2017a; Specia et al., 2021b), i.e., requiring labeled data for training, which is costly to obtain.

---

[1]Implementation available at `https://github.com/TuAnh23/Perturbation-basedQE`.

Additionally, supervised QE is likely to be domain-dependent and MT-system-dependent, as they do not aware of errors not occurring in the training data (Kocyigit et al., 2022).

Unsupervised QE overcomes the need for labeled data. Several works perform unsupervised QE by using glass-box features from the MT systems (Popović, 2012; Moreau and Vogel, 2012; Etchegoyhen et al., 2018; Niehues and Pham, 2019; Fomicheva et al., 2020b). As an example, Fomicheva et al. (2020b) proposed unsupervised QE using the output probability distribution and the attention mechanism from encoder-decoder MT models. Therefore, their methods are model-specific. Tuan et al. (2021) excludes the need for human-labeled data and MT glass-box access by creating synthetic data to train QE models. The synthetic data is generated by aligning candidate MT translations to the target references to find errors, or rewriting target reference sentences using a masked language model to introduce errors. These methods require a large and diverse amount of parallel MT data (i.e., source sentences and the gold-standard translations), which is not always available for different domains and language pairs. Additionally, these methods are also likely to be domain-dependent and MT-system-dependent, as the QE model is trained on the output of pre-selected MT systems on pre-selected MT data.

Researchers also focus on Quality Estimation from the explainability perspective. He et al. (2019) propose using integrated gradients from MT models (i.e., glass-box information) to quantify how important each source word is to the output translated words. The method is then used for QE by detecting under-translated source words that have low importance to the output translation. Ferrando et al. (2022) also quantify the contribution of each source word on the output translation using glass-box information from transformer-based MT models, which is the layer-wise tokens attributions. Here the source words' contribution can also be used to detect under-translated source words, or to assess the quality of the whole translation. Another line of research is on explainable sentence-level QE, where the word-level error scores are provided as the explanation for the predicted sentence-level score (Fomicheva et al., 2021). Explanations can be extracted by using methods such as LIME (Ribeiro et al., 2016) or SHAP (Lundberg and Lee, 2017) on top of sentence-level QE models, or building interpretable models that output both sentence-level quality and word-level explanations (Fomicheva et al., 2021).

In contrast to the previous works on Quality Estimation, Perturbation-based QE does not require labeled QE data, parallel MT data, nor glass-box access to the evaluated MT system. From the explainability perspective, our approach provides a new type of explanation for target-side word-level QE, i.e., the information on which source words affect each translated word.

3 Perturbation-based Quality Estimation

In this section, we describe Perturbation-based QE. Recall our motivation: if the MT system relies on too many tokens in the source sentence to output a translation token, it is likely that the system is exploiting irrelevant correlations, thus the translation token is unreliable.

Perturbation generation (Step I Figure 1): We first perform perturbation to the source sentence. The subset of source words to perturb, which is a hyperparameter choice, is one of the following: (1) content words, i.e., noun, verb, adjective, adverb, pronoun, determined by NLTK part-of-speech tagging (Bird et al., 2009); (2) all words including functional words such as "a", "an", "the"; or (3) all tokens including non-word tokens such as punctuation marks. For each perturbed source word s_i, we mask it out from the source sentence and use a language model to generate n best replacements. The language masking model can be BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) or DistilBERT (Sanh et al., 2019) (choice of the language masking model is a hyperparameter).

Translation (Step II Figure 1): We use the MT system to translate all perturbed versions.

Alignment (Step III Figure 1): We align at word level all the perturbed translations with the original translation. Two possible alignment methods are (1) Levenshtein (Levenshtein et al.,
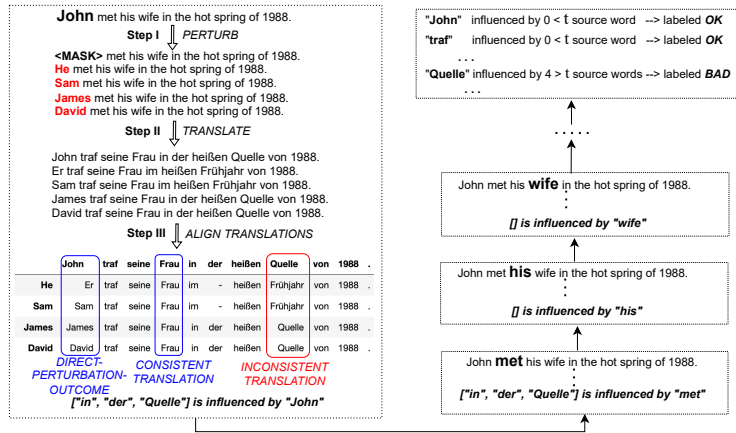
Figure 1: Perturbation-based QE. Words in the source sentence are perturbed one by one to find out their influence on the output words. If an output word h_j is influenced by more than t source words (excluding the source word directly translated to h_j), it is predicted as a BAD translation.

1966), which is the standard edit-distance alignment method that minimizes the number of insertion, deletion and substitution operations; and (2) Tercom (Snover et al., 2006), which additionally considers the shift operation. In Figure 1, the alignment outcome is shown in a table, where the column titles are the tokenized original translation, the row titles are the replacements of the perturbed source word, and each row is the aligned translation of the perturbed source sentence. Note that sometimes the alignments are not one-to-one. Some words in the original translation could have no aligned version in the perturbed translation. In this case, we align the original word with an empty token. Similarly, some words in the perturbed translation might not be aligned with any word in the original translation. In this case, we discard the words in the perturbed translation, since we only evaluate the consistency of the original words.

Consistency evaluation (Step III Figure 1): An MT-output word h_j is considered either a consistent translation, an inconsistent translation, or a direct-perturbation-outcome w.r.t. each perturbed source word s_i.

- Consistent translation is a translation that remains the same across perturbations. For example, in Figure 1, *"Frau"* is a consistent translation w.r.t perturbing *"John"*. To account for possible noise in alignment, we mark a translation as consistent if it remains the same across more than $c\%$ out of n perturbations w.r.t s_i. The threshold c is a hyperparameter.

- Direct-perturbation-outcome is the translation of the perturbed word, thus should vary in all perturbations. For example, in Figure 1, *"John"* is a direct-perturbation-outcome translation w.r.t perturbing *"John"*. To account for possible noise in translation and alignment, we mark a translation as direct-perturbation-outcome if the number of unique versions over the total n perturbations is larger than $p\%$. The threshold p is a hyperparameter.

- Inconsistent translation is a translation that has a few versions of it across n perturbations (i.e., the remaining cases). In Figure 1, *"Quelle"* is an inconsistent translation w.r.t perturbing *"John"*. When an MT-output word h_j is inconsistent due to perturbing a source word s_i, we say that h_j is influenced by s_i. Here *"Quelle"* is influenced by *"John"*.

Quality label prediction (Last block Figure 1): If the number of source words influencing h_j (excluding the one directly translated to h_j) is higher than a threshold t, then h_j is predicted

as a BAD translation, otherwise predicted as OK[2]. The threshold t is a hyperparameter.

Our approach comes with several advantages. First, it is unsupervised. The method does not rely on any labeled QE data or parallel MT data for training. This potentially makes the approach domain-independent and MT-system-independent. In other words, the approach would be robust to discover errors not presented in previous datasets, such as errors from a new MT system on a different domain. A small amount of labeled QE data can be used for hyperparameter tuning. However, our experiments show that the approach is not sensitive to hyperparameter choices, and that hyperparameters can be transferred across languages. Second, our approach is MT-system-agnostic and works for blackbox MT systems, as it only uses the MT system to generate translations. Third, our approach comes with explainability power. For each MT output word, our method shows which source words affect the generation of the considered output word. In this way, one can find wrong correlations inherent in the MT systems.

In terms of computational cost, Perturbation-based QE does not involve any training process. However, it requires computational power when using the evaluated MT system to generate translations of different perturbed versions of the source sentence. This can be considered as the trade-off between our approach and the previous QE approaches.

4 Experimental setup

4.1 Overall evaluation

Dataset: We use the word-level part of the MLQE-PE dataset (Fomicheva et al., 2020a), which is the benchmark in the WMT21 QE shared task (Specia et al., 2021b). The dataset consists of source sentences, the machine translation output and the word-level *OK/BAD* labels. We conduct experiments on four language pairs: English-German (*en-de*), English-Chinese(*en-zh*), English-Japanese (*en-ja*) and English-Czech (*en-cs*). In this dataset, *en-de* and *en-zh* directions are supervised, while *en-ja* and *en-cs* directions are zero-shot. However, we only use the development split for *en-de* and *en-zh* to perform hyperparameter tuning.

Evaluated MT systems: We use the to-be-evaluated encoder-decoder MT systems from WMT21, i.e., the fairseq Transformer (Ott et al., 2019) bilingual models for *en-de* and *en-zh*; and the ML50 fairseq multilingual Transformer model (Tang et al., 2020) for *en-cs* and *en-ja*.

Metrics: Following the WMT21 shared task, we use the Matthews correlation coefficient (MCC) (Matthews, 1975) as the evaluation metric for word-level QE in our experiments.

Hyperparameters: Hyperparameters for our approach, as explained in Section 3, are the number of unmasking replacements n, thresholds c, p, t, choices of source word subset for perturbation, language masking models to generate perturbation replacements and alignment tools. We use grid search to find the hyperparameter setting that yields the highest MCC score on development data. The best setting for *en-de* (which is then applied on *en-cs*) is $n = 30$, $c = 0.95$, $p = 0.9$, $t = 2$, perturbing content words, Tercom alignment and RoBERTa unmasking. The best setting for *en-zh* (which is then applied on *en-ja*) is $n = 30$, $c = 0.95$, $p = 0.8$, $t = 4$, perturbing all tokens, Tercom alignment and RoBERTa unmasking. We transfer the hyperparameters across languages in such a way since we expect more language-similarity between *de/cs* (alphabetic writing systems) and *zh/ja* (logographic writing systems).

Unsupervised QE baselines: For unsupervised baseline, we use the word-level log probabilities generated by the MT system. If the log probability of an output word is larger than a certain threshold l, then it is marked as OK, otherwise it is marked as BAD. The threshold l is a hyperparameter. Here we also use the development split to find the best l for *en-de* and *en-zh*. We apply the best value of l for *en-de* (which is $log_2 0.45$) on *en-cs* and the best value of l for

---

[2]We focus on evaluating the translation words that were output by the MT system. Our approach is not suitable to evaluate the gap between words or to detect untranslated parts of the source sentence.

en-zh (which is $log_2 0.60$) on *en-ja*. We choose this baseline since it has the same data usage as our approach, and it requires little information from the MT system (although here we no longer treat the MT system completely as blackbox).

Supervised QE baselines: We use the supervised baseline from the WMT21 QE shared task (Specia et al., 2021a). The baseline is a multilingual transformer-based Predictor-Estimator (Kim et al., 2017b), trained on labeled data for all available seven language directions. The model is trained multi-tasked, requiring both word-level and sentence-level labeled data.

4.2 Out-of-domain, unseen-MT-system evaluation

Common in-domain, known-MT-system setup: The common evaluation setup for QE approaches, e.g., in the WMT21 shared task, are in-domain and on known MT systems. That is, the QE test data is generated in the same way as the QE training data, and the to-be-evaluated MT system is the same as the one used to create the QE training data. However, in order to be useful in real-world applications, QE approaches should be capable of out-of-domain evaluation on unseen MT systems. That is, QE approaches should be able to evaluate different types of MT systems on different types of datasets. Therefore, we design experiments using QE approaches in an out-of-domain, unknown-MT-system setting, described as follows.

Evaluated MT systems: We test the QE approaches on evaluating two MT systems, one known and one unseen. The known MT system is the one that was used to create the WMT21 QE training and test data: the Fairseq encoder-decoder MT model. The unseen system is Flan-UL2 (available on HuggingFace) - a recent prompt-based large language model (LLM). We generate MT output from this LLM by prompting the system with "*Translate this into German: <English_input>*.". We choose this system as LLMs have been gaining a lot of attention and are more and more widely used (Vilar et al., 2022; Zhang et al., 2023; Bawden and Yvon, 2023). Going beyond the conventional encoder-decoder MT systems, we attempt to show that our approach is applicable to prompt-based translation using these prominent decoder-only LLMs.

Out-of-domain test data: We use two challenge sets on *en-de*. The first one is WinoMT (Stanovsky et al., 2019), used to evaluate gender bias from MT systems. WinoMT contains English input sentences with marked gender roles (e.g., "The **doctor** asked the nurse to help her in the operation") and evaluation protocol to identify whether the MT system outputs the correct gender form. The second challenge set is MuCoW (WMT 2019 translation test suite version) (Raganato et al., 2019), used to evaluate word-sense-disambiguation ability of MT systems. MuCoW contains English input sentences with ambiguous words and evaluation protocol to identify whether the MT system outputs the correct sense translations of the ambiguous words.

On WinoMT, the correct-gender accuracy is 69.4% for the Fairseq encoder-decoder MT system and 47.5% for the Prompt-based LLM Flan-UL2 system. On MuCoW, the correct-disambiguation accuracy is 47.59% for the Fairseq encoder-decoder MT system and 22.95% for the Prompt-based LLM Flan-UL2 system. Both MT systems do not perform well in outputting the correct gender form nor outputting the correct sense for ambiguous words. Therefore, it would be interesting to see whether QE methods can detect these mistakes.

Out-of-domain error detection: We test whether QE approaches can detect gender errors (which we refer to as GenderBAD tokens) and word-sense-disambiguation errors (which we refer to as WSD-BAD tokens). Given an MT system, we first generate translations for the WinoMT/MuCoW English sentences. Then we run WinoMT/MuCoW evaluation protocol to mark the GenderBAD/WSD-BAD tokens. An ideal QE approach should be able to detect all the GenderBAD/WSD-BAD tokens, i.e., correctly labeling them as BAD translations.

Metrics: We report on the GenderBAD-recall and WSD-BAD-recall, i.e., the percentage of GenderBAD/WSD-BAD tokens (marked by WinoMT/MuCoW) that are successfully predicted as BAD by the QE methods. We do not report on the GenderBAD/WSD-BAD accuracy,

since tokens with correct gender form or correct disambiguated sense are not necessarily OK translations. They could contain some other types of errors such as tense or singular/plural forms. Note that the recall metric could favor QE methods that are overly harsh (e.g., predicting everything as BAD would result in perfect GenderBAD/WSD-BAD recall). Therefore, we additionally report on the GenderBAD-precision and WSD-BAD-precision. Precision scores only serve as an indication of whether a QE model is too harsh. The reason is that, GenderBAD and WSD-BAD are not the only types of BAD error, thus it is not correct to always punish the QE model for predicting a non-GenderBAD or non-WSD-BAD token as BAD.

4.3 Robustness w.r.t hyperparameter choices

Recall that, in the previous experiments, we use the WMT21 development data of *en-de* and *en-zh* to perform hyperparameter tuning for Perturbation-based QE. The best hyperparameters for *en-de* are then also used for *en-cs*, and the ones for *en-zh* are used for *en-ja*, since we assumed more language similarity between *de/cs* and *zh/ja*. We refer to this as *"ideal hyperparameters"*.

The aim of this experiment is to test the robustness of our approach w.r.t hyperparameter choices, i.e., how much our approach relies on labeled development data. Similar to the experiment in Section 4.1, we report the MCC scores on in-domain setting, i.e., WMT21 test data. Here we apply the hyperparameter settings in an opposite way compared to previous experiment, i.e., (1) applying the best hyperparameter setting of *en-de* on *en-zh* and *en-ja*, and (2) applying the best hyperparameter setting of *en-zh* on *en-de* and *en-cs*. We refer to this as *"suboptimal hyperparameters"*. The MCC scores reducing significantly would indicate that our approach is sensitive to hyperparameters and vice versa. Additionally, we provide ablation experiments on the discrete hyperparameters to see their effects on the QE performance.

5 Results and Discussion

5.1 Overall QE performance

	Supervised/Tuned		Zero-shot/Un-tuned	
	en-de	en-zh	en-ja	en-cs
Log probability QE	0.241	0.149	0.112	0.257
WMT21 QE baseline	**0.370**	**0.247**	0.131	0.273
Perturbation-based QE	0.287	0.180	**0.218**	0.270

Table 1: Performance (in MCC score) of word-level QE approaches on WMT21 test data.

The performance of our Perturbation-based QE on the WMT21 test data, in comparison to other baselines, is shown in Table 1. Our approach outperforms the log-probability baseline over all language pairs. The largest gap is on *en-ja*, where our approach obtains 0.106 points higher. For *en-de* and *en-zh*, the gain from our approach is around 0.040 points. The smallest gain is on *en-cs*, where our approach outperforms the log-probability baseline by 0.013 points.

Compared to the WMT21 QE baseline, on *en-de* and *en-zh*, our approach falls behind by 0.083 and 0.067 points respectively. Recall that for these language pairs, the WMT21 QE is supervised, requiring labeled word-level data, and additionally uses sentence-level data for multi-task training. In contrast, our approach only uses the development split of word-level data for hyperparameter tuning. It can be concluded that our approach is not competitive to the supervised approaches that make use of more labeled data for training.

On *en-ja* and *en-cs*, our approach is competitive to the WMT21 QE baseline. Our approach outperforms the baseline by 0.087 points on *en-ja*, while having similar performance on *en-cs*. Recall that on these language pairs, both approaches do not use any labeled data of the language

pairs in consideration. The WMT21 baseline is zero-shot: it uses labeled training data of 7 other language pairs. In contrast, our approach only uses the development split of 2 other language pairs for hyperparameter tuning. It can be concluded that, when there is no direct labeled data for the language pair of interest, our approach has competitive performance while being more data efficient compared to the zero-shot approach. Additionally, observe that the performance of the WMT21 zero-shot baseline on *en-ja* is low compared to *en-cs*. This is possibly due to the low similarity between *ja* and other languages in the training data. This indicates that the zero-shot approach is more data-dependent, while this is not an issue for our approach.

5.2 Out-of-domain, unseen-system evaluation

The QE approaches' performance on detecting out-of-domain errors from known/unseen evaluated MT-system is shown in Table 2. As can be seen, our Perturbation-based QE approach has the best performance. When the task is to detect gender errors, our GenderBAD-recall is significantly higher than the second-best QE approach by over 0.200 points. When the task is to detect word sense disambiguation errors, our WSD-BAD-recall is higher than the second-best QE approach by over 0.049 points. At the same time, our GenderBAD-precision and WSD-BAD-precision scores are similar to the other methods, indicating that we are not overly predicting tokens as BAD to cheat for a higher GenderBAD-recall and WSD-BAD-recall.

	Known MT system (Encoder-Decoder MT)		Unseen MT system (Prompt-based LLM)	
	GenderBAD recall	GenderBAD precision	GenderBAD recall	GenderBAD precision
Log probability QE	0.175	0.036	0.429	0.047
WMT21 QE (supervised)	0.021	0.031	0.065	0.036
Perturbation-based QE	**0.391**	0.045	**0.658**	0.042
	Known MT system (Encoder-Decoder MT)		Unseen MT system (Prompt-based LLM)	
	WSD-BAD recall	WSD-BAD precision	WSD-BAD recall	WSD-BAD precision
Log probability QE	0.137	0.007	0.347	0.005
WMT21 QE (supervised)	0.290	0.028	0.177	0.004
Perturbation-based QE	**0.339**	0.009	**0.709**	0.005

Table 2: Results on detecting out-of-domain errors by different QE methods. The top half indicates results on WinoMT. The bottom half indicates results on MuCoW.

Another observation is that, the supervised WMT21 QE model performs poorly on detecting GenderBAD tokens from MT outputs on WinoMT. Its GenderBAD-recall is very low, at 0.021 for known MT system and at 0.065 for unseen MT system. This performance is even worse than that of the naive Log probability QE, whose GenderBAD-recall is 0.175 on known MT system output and 0.429 on unseen MT system output. A possible explanation is that, the WMT21 QE model does not aware of gender errors since this type of error does not present in the training data. This indicates the data-dependent issue inherent to supervised QE approaches.

The performance gap between our Perturbation-based QE approach and the supervised WMT21 QE model is larger on the unseen MT system than the known MT system. On the unseen MT system, our GenderBAD-recall is higher than the WMT21 QE model by +0.593 points, which is larger than the corresponding gap of +0.370 when evaluating the known MT system. Similarly, the WSD-recall gap is +0.532 on the unseen MT system, which is larger

than the gap of +0.049 on the known MT system. Additionally, when detecting word sense disambiguation errors from the unseen MT system, the performance of the WMT21 QE model is once again worse than the naive Log probability QE approach, where its WSD-BAD-recall is lower by -0.170 points. This indicates the MT-system-dependent issue inherent to supervised QE approaches, where they are not able to perform as well on evaluating unseen MT systems.

Overall, the supervised QE models, while performing better in a similar setting as their training process, fail behind Perturbation-based QE in out-of-domain and unseen-system settings. This strengthens the domain-independent and system-independent power of our approach: it can better detect errors from a new MT system on a new domain usage.

5.3 Robustness w.r.t hyperparameter choices

The difference in performance when using different hyperparameter settings in our approach is shown in Table 3. As can be seen, the MCC scores only deviate by around 0.010 points when using ideal versus suboptimal hyperparameter values. The difference in performance when using different values for the discrete hyperparameters is shown in Table 4. We consider 3 hyperparameters, in the top-to-bottom order displayed in Table 4: sets of perturbed source words, unmasking models and alignment methods. It can be seen that, the MCC scores generally deviate by only around 0.010 points. Overall, different choices of hyperparameters do not significantly affect our QE performance. This shows that our approach is not sensitive to hyperparameter choices, which is useful since we are not dependent on labeled data for hyperparameter tuning.

	en-de	en-zh	en-ja	en-cs
Ideal hyperparameters	0.287	0.180	0.218	0.270
Suboptimal hyperparameters	0.274	0.167	0.206	0.284

Table 3: Perturbation-based QE performance in MCC using ideal/suboptimal hyperparameters.

	en-de		en-zh	
	Best val MCC	Test MCC	Best val MCC	Test MCC
content words	0.282 ± 0.004	0.287 ± 0.005	0.196 ± 0.003	0.169 ± 0.005
all words	0.267 ± 0.004	0.277 ± 0.003	0.196 ± 0.002	0.162 ± 0.004
all tokens	0.270 ± 0.002	0.269 ± 0.006	0.208 ± 0.003	0.177 ± 0.004
BERT large	0.272 ± 0.006	0.278 ± 0.011	0.198 ± 0.006	0.174 ± 0.006
BERT base	0.272 ± 0.006	0.279 ± 0.009	0.199 ± 0.005	0.166 ± 0.006
DistilBERT	0.269 ± 0.008	0.272 ± 0.008	0.198 ± 0.008	0.168 ± 0.009
RoBERTa	0.277 ± 0.007	0.278 ± 0.008	0.203 ± 0.007	0.169 ± 0.007
Levenshtein	0.272 ± 0.006	0.275 ± 0.008	0.199 ± 0.006	0.170 ± 0.007
Tercom	0.274 ± 0.009	0.280 ± 0.009	0.200 ± 0.007	0.168 ± 0.007

Table 4: Ablation experiments on discrete hyperparameter settings for Perturbation-based QE. The performance of a setting in a specific group is averaged over all settings of the other groups.

5.4 Perturbation-based QE for explainable MT

We investigate some examples of using Perturbation-based QE for explaining the MT output gender errors on WinoMT and the word sense disambiguation errors on MuCoW. One gender error example is shown in Figure 2a. The MT system outputs the female form for *"housekeeper"*, while the form should be male, indicated by the word *"he"*. In an ideal scenario,

the gender form of *"housekeeper"* (*"Haushälterin/Haushalter"*) should only depend on *"he"*. However, our approach shows that, when perturbing the source word *"he"*, the output word *"Haushälterin"* does not change. Instead, when perturbing *["chief", "gave", "tip", "was", "helpful"]*, the output varies between *"Haushälterin"* and *"Haushalter"*, showing that the MT model is focusing on the wrong part of the sentence to determine the gender form.
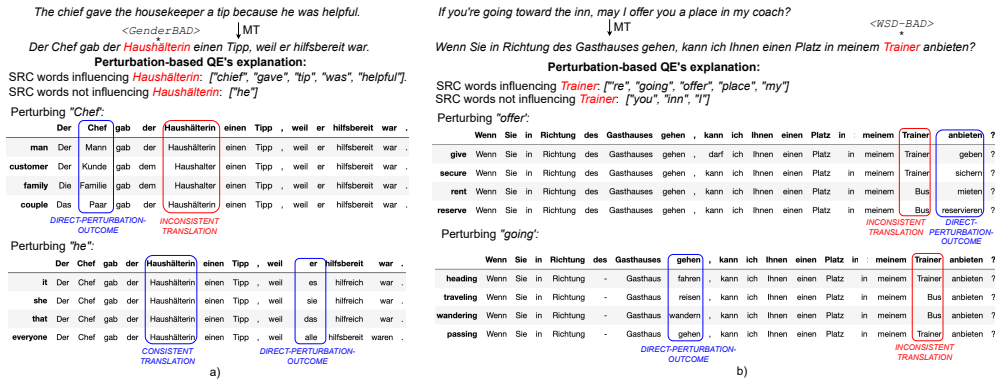

Figure 2: Example of Perturbation-based QE's explanation on WinoMT (a) and MuCoW (b).

One word sense disambiguation error example is shown in Figure 2b. The MT system outputs the wrong sense for *"coach"*. It outputs *"Trainer"*, which means the sports trainer. However, given the context, *"coach"* should mean the vehicle, thus the correct output should be *"Bus"*. Ideally, the MT system should only look at the context source words indicating movements to decide on the sense of *"coach"*. Nevertheless, our approach shows that the MT output translation for *"coach"* varies when perturbing multiple source words. For example, replacing *"offer"* with *"rent"* or *"reserve"* makes the system outputs the correct sense *"Bus"*, while for other replacements it still outputs *"Trainer"*. Similarly, when replacing *"going"* with other words that indicate movements, only *"traveling"* and *"wandering"* make the system outputs the correct sense *"Bus"*. This explanation provides an insight into the MT system: the sense *"Bus"* is only correlated to a few context words. Therefore, when the source sentence does not contain those specific words, the MT system fails to output the correct sense.

6 Conclusion

We proposed an unsupervised word-level Quality Estimation method, termed Perturbation-based QE. Our method does not rely on labeled QE data nor parallel MT data, masking it more domain-independent and system-independent to find MT errors that cannot be foreseen. This advantage is supported by our experiment on finding gender bias and word-sense-disambiguation erroneous translation from a nontraditional translation-prompting LLM. Our approach is not sensitive to hyperparameter settings, thus less dependent on labeled data for hyperparameter tuning. Our approach is also explainable: it shows which source words affect an output translated word. Additionally, our approach is MT-system-agnostic and works for black-box systems. Overall, Perturbation-based QE, as an unsupervised method, still falls behind supervised QE on in-domain and known-MT-system settings, but outperforms supervised QE on zero-shot settings and on out-domain and unseen-MT-system settings. As future work, it can be extended to assess the quality of other tasks, such as question answering or summarization, in the same manner: minimally perturb the input and analyze changes in the output.

References

Bawden, R. and Yvon, F. (2023). Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emelin, D., Titov, I., and Sennrich, R. (2020). Detecting word sense disambiguation biases in machine translation for model-agnostic adversarial attacks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7635–7653, Online. Association for Computational Linguistics.

Etchegoyhen, T., Martínez Garcia, E., and Azpeitia, A. (2018). Supervised and unsupervised minimalist quality estimators: Vicomtech's participation in the WMT 2018 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 782–787, Belgium, Brussels. Association for Computational Linguistics.

Ferrando, J., Gállego, G. I., Alastruey, B., Escolano, C., and Costa-jussà, M. R. (2022). Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fomicheva, M., Lertvittayakumjorn, P., Zhao, W., Eger, S., and Gao, Y. (2021). The Eval4NLP shared task on explainable quality estimation: Overview and results. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fomicheva, M., Sun, S., Fonseca, E., Zerva, C., Blain, F., Chaudhary, V., Guzmán, F., Lopatina, N., Specia, L., and Martins, A. F. (2020a). Mlqe-pe: A multilingual quality estimation and post-editing dataset. *arXiv preprint arXiv:2010.04480*.

Fomicheva, M., Sun, S., Yankovskaya, L., Blain, F., Guzmán, F., Fishel, M., Aletras, N., Chaudhary, V., and Specia, L. (2020b). Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

He, S., Tu, Z., Wang, X., Wang, L., Lyu, M., and Shi, S. (2019). Towards understanding neural machine translation with word importance. 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 953–962, Hong Kong, China. Association for Computational Linguistics.

Kim, H., Lee, J.-H., and Na, S.-H. (2017a). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568.

Kim, H., Lee, J.-H., and Na, S.-H. (2017b). Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 562–568, Copenhagen, Denmark. Association for Computational Linguistics.

Kocyigit, M., Lee, J., and Wijaya, D. (2022). Better quality estimation for low resource corpus mining. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 533–543, Dublin, Ireland. Association for Computational Linguistics.

Levenshtein, V. I. et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Moreau, E. and Vogel, C. (2012). Quality estimation: an experimental study using unsupervised similarity measures. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 120–126, Montréal, Canada. Association for Computational Linguistics.

Niehues, J. and Pham, N.-Q. (2019). Modeling confidence in sequence-to-sequence models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 575–583, Tokyo, Japan. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Popović, M. (2012). Morpheme- and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 133–137, Montréal, Canada. Association for Computational Linguistics.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2019). The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J. G., Glushkova, T., Alves, D. M., Lavie, A., et al. (2022). Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *arXiv preprint arXiv:2209.06243*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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.

Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., and Martins, A. F. (2021a). Findings of the wmt 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725.

Specia, L., Blain, F., Fomicheva, M., Zerva, C., Li, Z., Chaudhary, V., and Martins, A. F. T. (2021b). Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Tuan, Y.-L., El-Kishky, A., Renduchintala, A., Chaudhary, V., Guzmán, F., and Specia, L. (2021). Quality estimation without human-labeled data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 619–625, Online. Association for Computational Linguistics.

Vieira, L. N., O'Hagan, M., and O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2022). Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Zhang, B., Haddow, B., and Birch, A. (2023). Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.