There's no Data Like Better Data: Using QE Metrics for MT Data Filtering

Jan-Thorsten Peter* **David Vilar*** **Daniel Deutsch**
Mara Finkelstein **Juraj Juraska** **Markus Freitag**
Google
{jtp,vilar}@google.com

Abstract

Quality Estimation (QE), the evaluation of machine translation output without the need of explicit references, has seen big improvements in the last years with the use of neural metrics. In this paper we analyze the viability of using QE metrics for filtering out bad quality sentence pairs in the training data of neural machine translation systems (NMT). While most corpus filtering methods are focused on detecting noisy examples in collections of texts, usually huge amounts of web crawled data, QE models are trained to discriminate more fine-grained quality differences. We show that by selecting the highest quality sentence pairs in the training data, we can improve translation quality while reducing the training size by half. We also provide a detailed analysis of the filtering results, which highlights the differences between both approaches.

1 Introduction

In the times of statistical machine translation, a well-known motto was "there's no data like more data". Experimental results seemed to confirm this, with the performance of the systems steadily improving as more data was made available. Web crawling has proven to be a valuable source of data for training translation systems, with projects like Common Crawl[1] or ParaCrawl (Bañón et al., 2020) providing numerous parallel sentences with which to train MT systems. Inevitably, when crawling huge amounts of data, noise will be present. Taking the web as an example, the quality of available texts varies greatly between websites. There are sources which reliably produce high-quality text, e.g. large circulations newspaper websites usually contain text written and proof-read by professional journalists. But by its open nature, the web also contains texts of dubious quality (both in style and in content) which may pollute the collected texts.

When considering bilingual data collection, an additional difficulty comes into play, namely the alignment of segments between two or more languages. Sentence alignment algorithms (Gale and Church, 1993; Moore, 2002; Sennrich and Volk, 2011; Thompson and Koehn, 2019) are bound to make mistakes, resulting in pairing of sentences that are not necessarily translations of each other. Even if the correspondence between sentences may be correct, the quality of sentences may differ greatly between languages. While one source may provide high quality text in its original language, the available translations may be of sub-par quality due to a variety of reasons (Freitag et al., 2022b). In addition, given the increased availability (and quality) of machine translation engines, MT output is expected to be part of the crawled data, thus contaminating the training material.

Statistical systems were robust against such type of noise (Goutte et al., 2012). The maximum likelihood estimators of phrase probabilities (and related models) were based on relative frequencies, with the consequence that noisy translation units, while being available to the system at translation time, had a low chance of being used. In fact, works on filtering data dealing with statistical systems, e.g. Johnson et al. (2007) were more concerned with the efficiency of the systems, rather than with quality.

With the advent of neural machine translation, the situation has changed, and the quality of the data has a major impact in the resulting quality of the translation system. Neural networks have a great ability of memorizing (parts of) the training data (Arpit et al., 2017; Feldman and Zhang, 2020). Whereas for phrase-based models the noise was diluted in the abundance of better-quality data, in neural models such outliers may have a critical effect on the output of the system. Therefore data filtering has become increasingly important, even spawning dedicated shared tasks (Koehn et al., 2018, 2019,

* Equal contribution.
[1] https://commoncrawl.org/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 561–577
December 6–7, 2023.

2020). Research in this area has allowed NMT systems to take advantage of big amounts of training data. As an example, initial versions of ParaCrawl degraded translation quality when adding them unfiltered to the training data of MT systems (Junczys-Dowmunt, 2018b; Schamper et al., 2018), whereas nowadays it is one of the main data sources used in WMT for the languages where it is available.

Thanks to these filtering techniques, huge amounts of parallel sentences are available for several languages, in some cases reaching up to hundreds of millions of sentences. But as noted above, not all texts are of the same quality. It is only natural to ask the question, once the training data reaches a size which is "big enough", if all the available text is useful for training NMT systems, or if the lower quality sentences are hindering the system. Note that most data filtering systems are focused in detecting noise or problems in the translation (e.g. under- or overtranslation). In very wide strokes, most systems answer the question "Is sentence A a translation of sentence B?", without looking (too much) into the quality.

Judging the quality of translations is the focus of the Quality Estimation (QE) field of machine translation. It can be considered an extension of machine translation evaluation, where references are not available. In the last years, the use of neural models has improved the results in this area dramatically. Would it then be possible to use quality estimation methods for filtering data and improve the quality of a neural machine translation system, which has been trained on already cleaned data? In other words, can we do a more fine-grained data selection beyond discarding "obvious" errors, focusing on selecting the best data that can be found in the training corpus? We explore these questions in this paper.

Our scientific contributions are:

- We show that neural QE metrics are effective methods for data filtering.
- We analyze the differences between the sentences filtered by the methods and find out that QE methods are more sensitive toward quality differences, being able to detect bad quality translations or fine grained translation errors (e.g. wrong named entities in a perfectly valid translation).
- We show that on the other hand QE filtering cannot account for some actual noise problems. Thus a "traditional" method for filtering the raw data coming from crawls is still needed as a first step.

2 Related Work

Already in the early days of the popularization of statistical methods for machine translation, the potential of mining data from the web was recognized by Resnik (1999). In contrast to other data sources at the time (e.g. Canadian Hansards, European Parliament Proceedings) which consisted largely of clean data, the necessity of including additional "quality assurance" steps were recognized in this work. As pointed out above, statistical systems were robust against noisy input data (Goutte et al., 2012), and as such, the topic of "corpus filtering" was mainly focused on selecting subsets of data closer to a given domain (Axelrod et al., 2011). Nevertheless Taghipour et al. (2011) shows that statistical systems may also benefit from careful curation of the training data.

The situation changed dramatically with the advent of neural machine translation, as such systems are much more sensitive to noisy input data (Khayrallah and Koehn, 2018). A clear reflection of this fact was the creation of a new dedicated shared task in the WMT yearly conference (Koehn et al., 2018, 2019, 2020) in the years 2018 to 2020.[2]

Junczys-Dowmunt (2018a) was the best performing system in the first edition of the filtering shared task, using a cross-entropy approach between two translation systems trained on clean data. In the next edition, the focus was moved towards low resource conditions. That year Chaudhary et al. (2019) presented the best performing system, using a system based on LASER embeddings (Schwenk and Douze, 2017).

The 2020 edition continued the focus on low-resource languages. At that evaluation, three were the best performing systems: Lu et al. (2020) and Lo and Joanis (2020) both use pre-trained multilingual models as a key component of their filtering systems. Esplà-Gomis et al. (2020) used an improved version of BICLEANER, their submission for the 2018 campaign (Sánchez-Cartagena et al., 2018). The authors further improved their system (Ramírez-Sánchez et al., 2020), including an extension using neural models (Zaragoza-Bernabeu et al., 2022). This latest version is considered state-

[2] In this year's WMT there is a new related shared task: "Parallel Data Curation".

of-the art and it is that which we take as baseline for comparing our method.

Of course, this is just but a very rough overview of the best performing systems in each evaluation campaign. We refer to the reports of each campaign for a more detailed overview of the methods explored each year. Bane et al. (2022) provide one more recent overview of data filtering methods. In this work, the authors sample 5M sentences from original training data and added 1M noise samples manually. They show that a two-stage approach can be beneficial for improving the quality of a translation system.

All of these methods have a common focus on detecting the type of noise that may originate from crawled data. This type of noise has been analyzed in Khayrallah and Koehn (2018), and Herold et al. (2022) build on their work and carry out a comparison of the efficiency of different filtering methods on various types of noise. Kreutzer et al. (2022) also provide an extensive analysis on the noise present in several widely used corpora.

Most similar to our work Carpuat et al. (2017) start with already clean data and analyze the effect of semantic divergence on translation quality. They are able to effectively select a subset of the training data and improve translation quality measured in BLEU. Bernier-Colborne and Lo (2019) use YiSi-2, also a quality estimation metric as a component in their corpus filtering system for the WMT 2019 shared task. Lo and Simard (2019) extend this idea by including BERT (word) alignments in the YiSi pipeline. We follow a conceptionally similar approach to these papers, using state-of-the-art QE metrics and provide a more in-depth comparison to other corpus filtering methods more oriented towards noise detection.

Quality estimation is again its own area of research, with dedicated shared tasks, e.g. (Zerva et al., 2022), that measure how well metrics can predict word- and sentence-level quality scores. In contrast to traditional MT evaluation, QE aims to assess the quality of the output texts without the use of a reference translations. The most successful QE metrics learn to jointly predict word- and sentence-level scores, like COMETKIWI (Kepler et al., 2019; Rei et al., 2022). Another possibility is to modify the input to a learned reference-based metrics like BLEURT (Sellam et al., 2020) or COMET (Rei et al., 2020) to use the source segment instead of a reference translation to predict sentence-level quality scores (Rei et al., 2021). We follow the latter approach and train a QE version of BLEURT that predicts sentence-level quality scores (see Section 3) that are used for data filtering.

3 From BLEURT to BLEURTQE

The QE metric that we propose for data filtering is a learned MT evaluation metric that is based on a BLEURT-style architecture (Sellam et al., 2020). BLEURT is a reference-based regression metric that is trained to predict a quality score for a hypothesis translation given a reference. The hypothesis and reference are concatenated together with a special token in between, then fed as input to the metric, which predicts a floating point quality score.

Our QE metric is a modification of the original model. To make it a QE metric, we pass the source segment as input to the metric instead of the reference. Then, we follow the winning submission to the WMT'22 Metrics Shared Task (Freitag et al., 2022a), MetricX, and use a modified version of the mT5 encoder-decoder language model (Xue et al., 2021) as our network architecture. Not that these is a multilingual model, so the same system can be used for a variety of languages. The source and hypothesis are passed as input to the encoder, and an arbitrary logit from the first step of the decoder is trained to predict the hypothesis quality score.

The QE metric is trained on the direct assessment quality judgments that were collected as part of the WMT Metrics Shared Task from 2015-2020 (Bojar et al., 2015, 2016, 2017; Specia et al., 2018, 2020; Fonseca et al., 2019) for all available language pairs. To (meta-)evaluate the metric we measure its correlation with ground-truth translation quality ratings using the benchmark MQM dataset from WMT'22 (Zerva et al., 2022) that includes 3 language pairs: en-de, zh-en, and en-ru. Since our metric is used to score individual segments and not systems, we report the segment-level correlation between our metrics' scores and the gold MQM scores using Pearson's r and Kendall's τ, shown in Table 1. The correlations are competitive to the top QE submissions to the WMT'22 Metrics Shared Task.

A (more refined) version of this metric has been submitted to this year's QE shared task (Juraska et al., 2023), and has been open sourced. We refer the reader to the system description for a more fine-grained discussion of the details of the metric.

Metric	en-de		en-ru		zh-en	
	r	τ	r	τ	r	τ
UniTE-src	0.40	0.29	0.39	0.34	0.40	0.43
COMETKIWI	0.43	0.29	0.39	0.36	0.51	0.36
BLEURTQE	0.38	0.29	0.41	0.39	0.38	0.35

Table 1: Segment-level Pearson's r and Kendall's τ on the WMT'22 MQM ratings for our QE metric, BLEURTQE and the top-performing metrics in the WMT'22 Metrics Shared Task, COMETKIWI (Rei et al., 2022), UniTE-src (Wan et al., 2022).

4 Experiments

We report experiments on three language pairs: English ↔ German, Japanese ↔ English and Chinese ↔ English. Our starting point is the full training data as provided by the WMT evaluation campaign. Corpus sizes can be found in Table 3. As can be seen in that table, we are working on a medium-to-large data condition, with the smallest language pair already having over 30M sentence pairs.

One thing to note is that these datasets have already undergone a cleaning process by the WMT organizers. I.e. a system trained on the entirety of this data is already able to obtain very good performance. In fact, many of the systems participating in the WMT evaluations take the available data as-is.

For each language pair we will consider different ways to reduce the size to 50% of their original size. This value was chosen in preliminary experiments on the English to German data, and it is comparable to previous work (Bane et al., 2022). Fixing the target size beforehand also allows a fair comparison between all the methods.

4.1 Filtering Approaches

We will consider three different filtering approaches for our experiments.

4.1.1 Random Selection

The most straightforward method to reduce the size of the training data is to just randomly select the desired amount of sentence pairs. We do not expect this method to perform well, but it constitutes the most direct baseline for data size reduction.

4.1.2 BICLEANER

As a representative for the "noise-detection" corpus filtering methods we chose to use BICLEANER AI.[3] This tool is an extension of the previous BICLEANER tool. The underlying method is based on a classifier that predicts if a sentence is a translation of another. BICLEANER AI substitutes the original classifier, based on handcrafted rules and extremely randomized trees, with a neural classifier based on XLM-RoBERTa. Zaragoza-Bernabeu et al. (2022) provide a detailed description of the tool and present an extensive experimental comparison showing state of the art results for filtering ParaCrawl.

It is also worth noting that BICLEANER is part of the pre-processing pipeline for generating the ParaCrawl dataset.

4.1.3 Quality Estimation for Filtering

For testing the performance of QE metrics for filtering we use two state-of-the-art metrics, COMETKIWI[4] (Kepler et al., 2019) and BLEURTQE[5] as described in Section 3. For each sentence pair in the training data, we compute the QE score for the translation from English into the foreign language. We use these scores for filtering for both translation directions, i.e. the resulting parallel data is the same for English → Foreign than from Foreign → English. We are aware that this may introduce a certain bias, as the performance of the QE metrics is not symmetrical. However scoring the full training data is a costly operation as we have to run big neural models on tens or hundreds of millions of sentence pairs. We still expect to see improvements even when using the wrong direction for data filtering. The only exception may be the backtranslated portion of the Chinese ↔ English dataset: As the starting data is Chinese, the filtering method may miss low quality backtranslations produced by an automatic system.

4.2 Experimental Setup

For all the filtering methods (except random selection), we compute the score of each sentence pair, and then select a threshold as to keep 50% of the original data. We then train an NMT system from scratch using the resulting training data sets.

Our translation system is a transformer-based encoder-decoder model based on PaxML[6], very similar to most of the systems participating in the WMT evaluation campaign. It consists of 6 encoder

[3] https://github.com/bitextor/bicleaner-ai

[4] https://unbabel.github.io/OpenKiwi

[5] The tool will be open-sourced with the publication of the shared task system description.

[6] https://github.com/google/paxml

and 6 decoder layers, a model dimension of 1024, hidden dimension of 8192 and 16 attention heads. GELUs with gated activation are used as activation functions. We use a 32k shared vocabulary for each language pair, and limit the maximal sentence length to 128 tokens. The model has a total of 551M parameters.

We removed all sentences which have more than 128 tokens, but did not perform any other filtering or preprocessing of the data. All models are trained until they converged and we selected the checkpoint with the best BLEURT score on the WMT 2022 test sets.

In the discussion of the results we focus on the evaluation using COMET22. Traditional metrics like BLEU and CHRF are consistently outperformed by neural metrics in the WMT metrics shared task (Freitag et al., 2022a), thus we favor the use of such new metrics. We chose COMET22 over BLEURT in order to avoid overfitting on this last metric, as our proposed BLEURTQE model is based on it, and it also guides the checkpoint selection. Nevertheless, BLEURT, BLEU and CHRF scores are given in Appendix B.2 and confirm the trends reported here.

4.3 Test Data

In order to test on a variety of domains we use test sets from the WMT and IWSLT evaluation campaigns. We use the WMT 2019 (where available) consisting of news data, and the WMT 2022 and WMT2023 test sets, which are composed of a mix of different domains each. Additionally we experiment on the IWSLT'21 test set, sourced from TED talks (Anastasopoulos et al., 2021), and the IWSLT'23 dev set[7], which is based on ACL talks presentations (Agarwal et al., 2023).

Following the training data settings, we also filtered the test sentences longer than 128 tokens. As the WMT 2023 test set includes paragraph level evaluation, its size is reduced for en → de from 557 segments to 404 and for de → en from 549 to 468. All other test sets are barely affected (see Table 6 in Appendix A).

4.4 Experimental Results

Translation results for the English ↔ German language pair are shown in Table 2a. For en → de we can see that randomly selecting data hurts performance by 1 point on the WMT23 test set. Using each of the other filtering methods we are able to improve performance over using the full training dataset. For BICLEANER the improvement is rather modest, around 0.4 points for most test corpora. Note however that BICLEANER was already applied to the ParaCrawl dataset, which constitutes a big portion of the available training data for this language pair. As such it is understandable, or even expected, that translation quality is not improved by applying it again. The QE metrics perform similarly to each other, with a slight advantage of BLEURTQE over COMETKIWI. Using BLEURTQE we are able to achieve an improvement of up to 1.7 points on the WMT23 test set.

The results for de → en majorly confirm the previous observations. The best results are again achieved in this case on the WMT'23 data, with an improvement of 1.3 points achieved by both QE methods. For the WMT'22, IWSLT'21 and IWSLT'23 test sets, the translation performance basically stays constant for all filtering methods.

Results on English ↔ Japanese, shown in Table 2b also show similar trends. In this case the biggest improvement comprises 2.3 points on the WMT'23 test set[8], obtained by BLEURTQE. However for the ja → en translation direction we find an outlier, where no filtering achieves improvements over the baseline on the IWSLT'23 data.

Lastly, Table 2c shows the results for the Chinese ↔ English language pair. Again we can confirm the same trends as for the other two language pairs. The QE metrics are able to improve up to 2.8 points for the WMT'23 test set. The IWSLT'23 dataset again fails to achieve improvements, and in this case BICLEANER deteriorates translation quality, while the QE metrics are able to keep the performance.

Overall, we see that the QE metrics are effective in improving translation quality while retaining just half of the training data. The improvements can rage up to more than 2 COMET22 points, depending on language pair and test set. For IWSLT'23, having a more specialized technical domain, the QE metrics are not able to improve quality, for several language directions. But except for the case of English → Japanese, they also do not hurt performance. Additional results differentiating between the single domains of the WMT'22 and WMT'23 corpora can be found in Appendix B.1. In Ap-

[7]We use the dev set for IWSLT'23, since the test set is currently not publicly available.

[8]A slightly bigger improvement of 2.4 is obtained for WMT'22, but we skip this as we used this corpus to choose the best checkpoint during training.

(a) COMET22 scores for en ↔ de experiments.

	Filter	WMT'22 (dev)	WMT'19	WMT'23	IWSLT'21	IWSLT'23
en → de	Random	84.0	85.5	80.8	82.8	84.2
	None	86.2	86.0	81.8	83.2	84.5
	BICLEANER	86.7*	86.4*	82.2	83.3	84.9
	COMETKIWI	86.9**	**86.7****	82.9*	83.6*	84.8
	BLEURTQE	**87.2*****	**86.7****	**83.5****	**83.9****	**85.1***
de → en	Random	84.2	84.3	83.5	84.1	87.2
	None	84.5	84.6	83.3	84.2	**87.4**
	BICLEANER	84.2*	84.8	84.0*	84.1	87.3
	COMETKIWI	84.6*	85.1*	**84.6****	**84.4***	87.3
	BLEURTQE	**84.8****	**85.2***	**84.6****	**84.4***	87.3

(b) COMET22 scores for en ↔ ja experiments.

	Filter	WMT'22 (dev)	WMT'23	IWSLT'23
en → ja	Random	84.5	80.7	85.8
	None	85.6	82.3	86.9
	BICLEANER	86.0*	83.2*	87.2
	COMETKIWI	86.6**	83.7**	**87.9***
	BLEURTQE	**87.0*****	**84.0*****	87.4
ja → en	Random	75.9	75.0	84.6
	None	77.6	75.9	**85.5***
	BICLEANER	78.1*	77.4*	85.0
	COMETKIWI	78.7**	78.0**	85.0
	BLEURTQE	**79.0*****	**78.2****	85.1

(c) COMET22 scores for en ↔ zh experiments.

	Filter	WMT'22 (dev)	WMT'19	WMT'23	IWSLT'23
en → zh	Random	80.2	77.2	79.3	82.1
	None	81.2	77.6	79.7	**84.2***
	BICLEANER	81.7*	78.4*	80.3*	83.3
	COMETKIWI	83.0**	**80.1****	**82.5*****	84.1*
	BLEURTQE	**83.4*****	79.9**	82.2**	84.1*
zh → en	Random	72.2	78.1	74.2	84.1
	None	72.8	78.3	74.7	**84.9***
	BICLEANER	74.8*	79.6*	75.7*	84.2
	COMETKIWI	75.2**	**80.0****	**76.0****	84.5
	BLEURTQE	**75.4****	79.9**	**76.0****	84.8*

Table 2: COMET22 scores all experiments. For each language direction, systems marked with stars are statistically significantly better than systems with fewer stars (pairwise permutation test (Koehn, 2004) with p=0.05). "Random" was excluded from the significance computation.

Language pair	Full	Filtered	Common
de ↔ en	292.8M	146M	105M
en ↔ zh	55.2M	27.6M	17.5M
en ↔ ja	33.9M	16.9M	10.4M

Table 3: Amount of sentences before filtering, after filtering, i.e. 50% of the original corpus size, and number of sentences kept by both BLEURTQE and BICLEANER. All language directions include ParaCrawl data. English ↔ Chinese includes around 19.7M backtranslated Chinese sentences, as provided by the WMT organizers.

pendix B.2 we also report the results of the same experiments using BLEU, CHRF and BLEURT. These metrics confirm the observations presented in this section.

Koehn et al. (2020) mentions that on average metrics that select shorter sentences performed better on Parallel Corpus Filtering. Contrary to that we observed that dropping 50% of the data with the proposed method led to on average slightly longer sentences e.g. from 14.4 to 15.5 words per sentence for BLEURTQE on English ↔ German.

5 Analysis

In this section we provide an in-depth analysis of the differences between the BLEURTQE-based and the BICLEANER filtering methods. Table 3 shows the amount of sentence pairs that are kept by both methods, which is roughly two thirds of the filtered sentences for all language pairs. Thus, it is clear that both methods do indeed perform quite different filtering. We will first report on manual inspection of the most striking divergences between both methods. In Section 5.2 we will then provide a more quantitative analysis of the behaviour of the methods using synthetic data.

5.1 Human Inspection

We will now analyze the difference in the filtering methods by looking into the sentences that are selected by each method. To this end, we select the sentences where one method filters it but the other does not. In addition we use automatic clustering methods in the spirit of (Aharoni and Goldberg, 2020) in order to get insights about topic distribution. We limit our analysis to the German–English language pair[9], but as the methods are largely language independent, we feel confident that our findings will generalize to the other language pairs. Also, due to the fuzzy and partially subjective nature of this investigation, we are unable to provide exact statistics about each kind of effect.[10]

We have encountered the following major differences in the working of the methods. For each of these categories it is easy to find an abundance of examples (easily in the thousands) in the filtered data.

Single Entity Mistranslations When looking into the parallel data available for training, one can find a big amount of "templated texts", i.e. sentences that have a common structure, but that differ in one or few components, frequently named entities or numbers. Some examples can be found in Table 4a. The first entry in this table is a typical example. In the travel domain, there is a big amount of sentences of the form "Flights from *cityA* to *cityB*", "Hotels in *city*" or similar formulations. One frequent source of sentence alignment errors originates from sentences that follow the same template, but have different instantiations. Although the travel domain is one of the biggest representative of these type of sentences, it is by no mean the only one, as the other examples in Table 4a show, including the financial and the technical domain.

In these type of sentences, the QE metric seems to be more sensitive to alignment errors. All the sentences shown in Table 4a (and many others) are selected by BICLEANER, while BLEURTQE discards them.

Low Quality Translations In this category we include training examples where one or both sides are of low quality. Examples can be found in Table 4b. We can see that the language quality of the examples is borderline at best. Strictly speaking, the translations are "correct" in the sense that they preserve the structure of the sentence. As such BICLEANER gives them a relative high score and are kept in the training corpus. BLEURTQE, on the other hand, is explicitly trained to flag such erroneous sentences (as they might very well originate from MT engines), and thus these examples are filtered out.

Bad Related Sentence Alignments As pointed out above, sentence alignment is also an automatic process. While both methods perform quite well when detecting clearly bad aligned sentences (see Section 5.2), we found that there are cases where

[9] None of the authors are speakers of Japanese or Chinese.

[10] If we were able to develop such statistics in an automatic way, we would be able to improve the filtering methods by including the same approaches!

(a) Single Entity Mistranslations. Templated texts where the specific instantiation is different in both languages. BLEURTQE filters out these examples, while BICLEANER keeps them.

English	German	Comments
Flights from Tallinn to Stockholm	Flüge ab Tallinn nach Friedrichshafen	"Stockholm"' changed to "Friedrichshafen".
Total EU spending in Germany – € 11.013 billion	Gesamtzuschüsse der EU in den Niederlanden: 2,359 Milliarden EUR	Land and amount changed.
Documents that we receive from a manufacturer of a Redball Electrical 565 can be divided into several groups.	Dokumente, die wir vom Produzenten des Geräts Trevi AVX 565 erhalten, können wir in mehrere Gruppen teilen.	Product code changed.

(b) Examples of low quality sentences in at least one of the languages. BLEURTQE filters out these examples, while BICLEANER keeps them.

English	German	Comments
We are both, we have own factory which can ensure sculpture quality and best price and have a profession team to provide you best service.	Wir sind beide, wir haben eigene Fabrik, die Skulpturqualität und besten Preis sichern kann und ein Berufsteam haben, um Ihnen besten Service zur Verfügung zu stellen.	Unnatural language on both sides.
We honor do not track signals and do not track, plant cookies, or use advertising when a Do Not Track (DNT) browser mechanism is in place.	Wir achten darauf, dass Sie keine Signale verfolgen und keine Cookies verfolgen oder Cookies verwenden, wenn Sie einen DNT-Browser-Mechanismus (Do not Track) verwenden.	Unnatural language on both sides.
It really is fast, easy, free and additionally to attempt.	Es ist schnell, Schnell, gratis und am besten von allen zu try.	Incorrect sentences in both languages.

(c) Examples of wrong sentence alignment, although the sentences are related to each other. BLEURTQE filters out these examples, while BICLEANER keeps them.

English	German	Translated German
Could you help me? Help to improve my English and French language	Ja, ich möchte gern mein Deutsch mit Dir verbessern.	Yes, I want to improve my German with you.
We, therefore, guarantee that you will get daily updates on office spaces to rent in Hong Kong.	Wir können Ihnen deshalb versichern, dass Sie bei uns täglich einen aktuellen Überblick über den österreichischen Markt erhalten.	We can guarantee that you will get an up-to-date daily overview about the Austrian market.
This implies that the law is either repealed or not enforced.	Darüber hinaus wird sichergestellt, dass bestehende Gesetze nicht dupliziert oder konterkariert werden.	In addition, it is ensured that existing laws are not duplicated or counteracted.

(d) Examples of sentence pairs originating from the Bible. BLEURTQE filters them out, probably due to archaic language, while BICLEANER keeps them.

English	German
19 Behold, my belly is as wine which hath no vent; it is ready to burst like new bottles.	19 Siehe, mein Bauch ist wie der Most, der zugestopfet ist, der die neuen Fässer zerreißet.
7:16 Those who went in, went in male and female of all flesh, as God commanded him; and Yahweh shut him in.	7:16 und das waren Männlein und Fräulein von allerlei Fleisch und gingen hinein, wie denn Gott ihm geboten hatte.

Table 4: Example sentences where the filtering methods diverge.

the source and target sides are related and the BICLEANER system seems to get confused by this proximity. Some examples are given in Table 4c. It can be seen that in all three examples the German side is clearly related to the English text, with probably a overlap big enough to get an acceptable score from the translation system underlying BICLEANER. Again, as BLEURTQE is trained to distinguish fine-grained differences between translations, it is more robust against this kind of problems.

Religious Texts One shortcoming we found for the BLEURTQE method is that many sentences originating from the Bible corpus (or similar religious texts) are filtered out, while BICLEANER keeps them. Some examples are given in Table 4d. This is probably due to the language being archaic, very different to the type of sentences BLEURTQE has been trained on. Such style would be heavily penalized in an evaluation, as a more modern language would be preferred.

5.2 Noise

In the previous section we saw several examples where BLEURTQE outperforms BICLEANER for data selection. However we should not forget that BICLEANER was developed with a (related but) different goal, namely the cleaning of raw data. In fact, our starting datasets, as made available for the WMT evaluation have already undergone a cleaning process, and are already at a pretty high quality level.

If we were dealing with crawled data directly, we would need to address different phenomena. In this section we study how the filtering methods perform when dealing with the typical noise found on crawled data. We follow Herold et al. (2022) for the categorization of different noise types, which in turn is based on Khayrallah and Koehn (2018). We create synthetic data for the English → German translation direction containing the following noise categories:

Misaligned Sentences created by shuffling the target side of the corpus.

Misordered Words created by reordering the words in either the source or the target sentences.

Wrong Language created by taking parallel sentences corresponding to another language pair.

Untranslated created by copying one sentence into the other direction i.e. each sentence pair in the corpus has a copy of the source sentence as a "target" sentence (or the reverse direction).

Over/Undertranslation created by truncating either the source or the target side.

We refer the reader to Herold et al. (2022) for a more detailed description and justification of these categories. We omitted the "Short Segments", "Raw Crawled Data" and "Synthetic Translations" categories, as it was not clear how to define the correct filtering strategy in those cases.

For each of the studied categories, we generated 200K synthetic noise examples by randomly selecting a subset of the training data. For these experiments we re-tuned the threshold for each method by computing the scores for the original sentences and the noise examples, and computing the median. In this way, we filter exactly half of the data and a perfect system would be able to completely separate the original examples from the noisy ones.

Results can be found in Table 5. It can be seen that for most categories BICLEANER clearly outperforms BLEURTQE. This is specially the case for the "Wrong Language" and "Untranslated" categories, where BICLEANER can detect all the noisy examples. In fact, one of the practical advantages of BLEURTQE is at the same time one of its weaknesses. As its backbone model is a multilingual model, it is able to handle a wide number of languages, but it does not have a way to differentiate between them.

For "Misordered Words" we find an interesting asymmetry. BLEURTQE is much stronger in detecting problems when the target side is reordered, undoubtedly due to this being the "natural" direction for which it was trained. BICLEANER also shows this behavior, with its target side performance being superior to that of BLEURTQE, but inferior in the opposite direction. BICLEANER is also clearly better at detecting Over- and Undertranslations.

5.3 Combination of Filtering Methods

Since BLEURTQE and BICLEANER based filtering both improve translation quality, and they filter different sentences, it is only natural to try to combine both. As can be seen in Table 3, the amount of available data dropped to roughly one third when combining both methods. The result

Noise type	BLEURTQE	BICLEANER
Misaligned Sentences	8.1	**5.6**
Misordered Words (src)	**24.3**	39.1
Misordered Words (tgt)	10.3	**6.3**
Wrong Language	43.8	**0.0**
Untranslated (src)	47.6	**0.0**
Untranslated (tgt)	63.8	**0.0**
Overtranslation	35.2	**13.9**
Undertranslation	13.4	**7.5**

Table 5: Percentage of sentences being kept as valid for each of the synthetic noise categories. 0% means that all noisy sentences have been filtered out, i.e. perfect performance.

only slightly degraded in the English to German direction compared to just using BLEURTQE (0.3 BLEURT points on WMT'22), but degraded more in the German to English direction (0.6 BLEURT points on WMT'22). Combining the two methods is thus too aggressive with our setup, and hurts translation performance. Adapting the thresholds for the combination, may result in better performance. Note however that ParaCrawl already used BICLEANER in its pipeline (Esplà et al., 2019), thus we have already implicitly been using a combination of both methods.

6 Conclusions

In this paper we have shown that filtering data using QE metrics is an effective way of improving translation quality. In contrast to "traditional" data filtering methods that focus on detecting noise in the data, QE methods focus on selecting the best translation examples. Analyzing the differences between the two different methods, we see that QE metrics are not as effective at detecting certain types of noise, e.g. untranslated sentences, but are much better at identifying more fine grained problems in the data, like small translation errors or grammatical mistakes. Therefore, when starting with already cleaned data, we can obtain a boost in performance by focusing the NMT system training on the best sentences.

Our results show that the improvements obtained generalize across different domains, as measured by a variety of metrics. Even for more distant domains, like the ACL Talks of of the IWSLT'23 corpus, the performance of the systems remains largely constant. QE estimation is a very active field of research. Using this approach, the improvements obtained in this area can have a direct impact on improving the quality of NMT systems.

Limitations

Better results could have been obtained by tuning the threshold for each method individually, but this would also increase the computational cost massively.

A more in-depth comparison could be carried out starting from the raw web-crawled data. However in this study we chose to start from conditions similar to what most participants in the WMT evaluation use.

Ethics Statement

BLEURTQE and COMETKIWI scoring all the training data is computationally expensive, and may be a limiting factor of the method for small institutions.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner.

2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Fred Bane, Celia Soler Uguet, Wiktor Stribiżew, and Anna Zaretskaya. 2022. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 313–325, Orlando, USA. Association for Machine Translation in the Americas.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Gabriel Bernier-Colborne and Chi-kiu Lo. 2019. NRC parallel corpus filtering system for WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 252–260, Florence, Italy. 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.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 952–958, Online. Association for Computational Linguistics.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891. Curran Associates, Inc.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi,

George Foster, Alon Lavie, and André F. T. Martins. 2022a. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022b. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, San Diego, California, USA. Association for Machine Translation in the Americas.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018a. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018b. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 425–430, Belgium, Brussels. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 Metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual

datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Chi-kiu Lo and Eric Joanis. 2020. Improving parallel data identification using iteratively refined sentence alignments and bilingual mappings of pre-trained language models. In *Proceedings of the Fifth Conference on Machine Translation*, pages 972–978, Online. Association for Computational Linguistics.

Chi-kiu Lo and Michel Simard. 2019. Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 206–215, Hong Kong, China. Association for Computational Linguistics.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 979–984, Online. Association for Computational Linguistics.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA. Springer.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Julian Schamper, Jan Rosendahl, Parnia Bahar, Yunsu Kim, Arne Nix, and Hermann Ney. 2018. The RWTH Aachen University supervised machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 496–503, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

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

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Appendices

A Test Data Statistics

Table 6 shows statistics of the test data sets after filtering the sentences with a length of over 128 tokens.

	test set	lines	filtered
WMT'22	en → de	2037	2036
	de → en	1984	1981
	en → ja	2037	2037
	ja → en	2008	2007
	en → zh	2037	2037
	zh → en	1875	1849
WMT'23	en → de	557	404
	de → en	549	468
	en → ja	2074	2073
	ja → en	1992	1988
	en → zh	2074	2073
	zh → en	1976	1948

Table 6: WMT test set sizes. All test sets are filtered to use less than 128 tokens. This mainly reduced the en ↔ de WMT'23 test set since this was a paragraph level task. The effect on all other test sets is minimal.

B Additional Results

B.1 WMT Data: Domain-specific Evaluation

The WMT'22 and WMT'23 test sets are comprised of text originating from different domains. The scores reported in the main text correspond to the evaluation of the corpora as a whole. In Tables 7 to 12 we show the results for each individual domain. It can be seen that the improvements are achieved over all separate domains. There are only two cases where training on all data performs slightly better than filtering with BLEURTQE (ecommerce de → en in Table 8, and manuals zh → en in Table 10).

B.2 Other Metrics

In this appendix we report the COMET22, BLEURT, BLEU and CHRF scores for the experiments reported in Section 4. Table 13 shows the results for German → English, Table 14 for English → German, Table 15 for English → Japanese, Table 16 for Japanese → English, Table 17 for English → Chinese and Table 18 for Chinese → English. The additional metrics support the conclusions of the paper.

	WMT'22				WMT'23			
	conversation	ecommerce	news	social	mastodon	news	speech	user review
None	87.9	87.4	86.4	83.1	82.5	80.8	80.9	82.0
BICLEANER	88.7	88.1	86.8	83.1	82.5	82.3	81.4	82.1
COMETKIWI	88.8	88.3	86.8	83.9	83.3	82.9	81.6	83.8
BLEURTQE	**88.9**	**88.5**	**87.2**	**84.2**	**84.2**	**83.1**	**81.8**	**84.0**

Table 7: COMET22 scores for each domain of en → de WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	84.7	**85.4**	84.4	83.5
BICLEANER	84.8	85.0	84.5	82.8
COMETKIWI	85.0	85.2	84.8	83.6
BLEURTQE	**85.1**	85.3	**84.9**	**83.9**

Table 8: COMET22 scores for each domain of de → en WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	84.2	84.0	81.5	75.3
BICLEANER	85.2	83.9	81.8	76.1
COMET22	86.0	84.8	83.3	78.1
BLEURT	**86.4**	**84.9**	**83.6**	**78.7**

Table 9: COMET22 scores for each domain of en → zh WMT test sets.

	WMT'22				WMT'23		
	conversation	ecommerce	news	social	manuals	news	user review
None	74.0	66.8	76.8	74.1	**77.7**	78.9	68.4
BICLEANER	75.2	70.8	77.9	75.6	77.4	79.5	70.5
COMETKIWI	**76.4**	71.2	78.2	75.7	77.5	**80.1**	70.6
BLEURTQE	75.9	**71.5**	**78.3**	**76.0**	77.5	79.7	**71.0**

Table 10: COMET22 scores for each domain of zh → en WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	88.1	87.5	86.4	80.6
BICLEANER	88.6	87.7	86.5	81.0
COMETKIWI	89.2	87.9	87.3	81.9
BLEURTQE	**89.3**	**88.6**	**87.7**	**82.5**

Table 11: COMET22 scores for each domain of en → ja WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	77.5	83.0	75.2	74.9
BICLEANER	77.0	83.1	77.1	75.2
COMETKIWI	77.4	**84.1**	77.8	75.4
BLEURTQE	**78.2**	83.9	**78.3**	**75.5**

Table 12: COMET22 scores for each domain of ja → en WMT test sets.

	WMT'22 (dev)				WMT'19				WMT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	84.5	73.4	32.3	57.4	84.6	73.1	41.1	64.8	83.3	72.2	37.4	61.2
Random	84.2	73.1	32.0	57.1	84.3	72.9	40.5	64.4	83.5	72.2	38.2	62.1
BICLEANER	84.2	73.0	32.0	57.2	84.8	73.4	41.3	65.5	84.0	73.1	38.9	63.5
COMETKIWI	84.6	73.6	**32.5**	**57.5**	85.1	**74.0**	**41.8**	**65.7**	**84.6**	73.8	**40.6**	**65.0**
BLEURTQE	**84.8**	**73.7**	32.3	57.4	**85.2**	**74.0**	41.4	65.3	**84.6**	**73.9**	39.9	64.2
	IWSLT'21				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	84.2	73.2	27.8	52.7	**87.4**	**79.1**	**47.2**	**72.4**
Random	84.1	73.0	27.3	52.5	87.2	78.5	45.6	71.2
BICLEANER	84.1	73.0	27.5	52.5	87.3	79.0	46.9	72.1
COMETKIWI	**84.4**	**73.3**	**28.0**	**53.0**	87.3	79.0	46.6	71.9
BLEURTQE	**84.4**	**73.3**	**28.0**	**53.0**	87.3	79.0	47.1	72.2

Table 13: Full results for German → English. The COMETKIWI and BLEURTQE results for IWSLT'21 are identical due to rounding.

	WMT'22 (dev)				WMT'19				WMT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	86.2	76.7	35.9	62.6	86.0	75.7	43.1	67.0	81.8	70.1	38.2	61.6
Random	86.0	76.5	35.1	62.2	85.5	75.0	42.4	66.5	80.8	68.9	36.4	61.0
BICLEANER	86.7	77.5	36.4	63.1	86.4	76.2	42.9	67.2	82.2	70.5	38.4	62.3
COMETKIWI	86.9	77.6	36.2	63.0	**86.7**	76.4	**44.0**	**68.0**	82.9	71.9	**40.9**	**65.9**
BLEURTQE	**87.2**	**78.0**	**36.7**	**63.3**	**86.7**	**76.5**	42.1	66.8	**83.5**	**72.4**	**40.9**	65.5
	IWSLT'21				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	83.2	73.0	23.2	56.6	84.5	76.5	45.8	72.2
Random	82.8	72.8	22.8	56.3	84.2	76.2	44.9	71.7
BICLEANER	83.3	73.1	23.1	56.7	84.9	**76.8**	45.2	71.9
COMETKIWI	83.6	73.5	23.2	56.9	84.8	76.6	**45.9**	**72.4**
BLEURTQE	**83.9**	**74.0**	**23.9**	**57.2**	**85.1**	**76.8**	45.6	72.0

Table 14: Full results for English → German.

	WMT'22 (dev)				WMT'23				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	85.6	64.5	22.2	31.4	82.3	58.3	19.0	28.9	86.9	68.1	40.5	48.1
Random	84.5	62.8	21.0	30.3	80.7	55.5	17.2	27.2	85.8	65.8	35.7	43.6
BICLEANER	86.0	64.9	21.9	31.5	83.2	59.2	19.1	29.0	87.2	68.2	39.2	47.1
COMETKIWI	86.6	65.5	**22.7**	32.1	83.7	**60.0**	19.3	**29.5**	**87.9**	**69.3**	**41.3**	**48.6**
BLEURTQE	**87.0**	**66.1**	**22.7**	**32.2**	**84.0**	**60.0**	**19.5**	**29.5**	87.4	68.1	38.8	46.5

Table 15: Full results for English → Japanese.

	WMT'22 (dev)				WMT'23				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	77.6	62.6	18.1	42.5	75.9	62.2	16.0	41.6	**85.5**	**73.4**	**30.5**	**62.2**
Random	75.9	60.8	16.1	40.8	75.0	61.0	15.0	40.7	84.6	71.9	27.4	59.9
BICLEANER	78.1	63.6	18.2	44.3	77.4	63.3	17.2	44.2	85.0	72.2	28.4	61.0
COMETKIWI	78.7	64.1	18.8	44.6	78.0	64.0	16.6	44.1	85.0	72.6	28.2	60.8
BLEURTQE	**79.0**	**64.4**	**19.0**	**45.2**	**78.2**	**64.3**	**17.4**	**44.7**	85.1	72.8	29.6	61.7

Table 16: Full results for Japanese → English.

	WMT'22 (dev)				WMT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	81.2	65.8	37.7	33.5	79.7	64.8	43.4	39.0
Random	80.2	64.5	36.6	32.7	79.3	64.3	41.6	37.3
BICLEANER	81.7	66.6	37.0	33.0	80.3	65.7	42.4	37.6
COMETKIWI	83.0	68.0	38.2	34.0	**82.5**	**68.2**	**44.0**	**40.1**
BLEURTQE	**83.4**	**68.5**	**38.7**	**34.4**	82.2	67.7	43.6	38.9
	WMT'19				IWSLT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	77.6	59.4	31.1	27.4	**84.2**	**72.2**	**52.5**	47.1
Random	77.2	59.1	30.7	27.5	82.1	69.3	47.6	41.9
BICLEANER	78.4	60.4	31.1	27.5	83.3	70.7	47.9	**42.3**
COMETKIWI	**80.1**	**62.2**	**32.1**	**28.3**	84.1	71.2	47.9	42.2
BLEURTQE	79.9	61.9	31.7	28.1	84.1	71.1	47.3	42.2

Table 17: Full results for English → Chinese.

	WMT'22 (dev)				WMT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	72.8	59.5	17.4	46.2	74.7	61.0	18.8	44.9
Random	72.2	58.8	16.9	46.3	74.2	60.2	18.5	44.9
BICLEANER	74.8	60.9	17.1	47.6	75.7	61.5	19.1	45.8
COMETKIWI	75.2	61.4	**17.9**	**48.5**	**76.0**	61.6	**19.2**	**46.2**
BLEURTQE	**75.4**	**61.7**	17.7	48.1	**76.0**	**61.9**	18.6	45.7
	WMT'19				IWSLT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	78.3	65.6	23.7	53.6	**84.9**	**74.5**	**33.3**	**63.2**
Random	78.1	65.0	23.2	53.2	84.1	73.8	31.7	62.1
BICLEANER	79.6	66.6	23.9	54.7	84.2	73.9	32.6	62.4
COMETKIWI	**80.0**	67.0	**24.7**	**55.5**	84.5	73.9	30.8	61.6
BLEURTQE	79.9	**67.2**	23.7	54.9	84.8	**74.5**	32.1	62.5

Table 18: Full results for Chinese → English.