Detecting Various Types of Noise for Neural Machine Translation

Christian Herold **Jan Rosendahl** **Joris Vanvinckenroye** **Hermann Ney**
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
{surname}@i6.informatik.rwth-aachen.de

Abstract

The filtering and/or selection of training data is one of the core aspects to be considered when building a strong machine translation system. In their influential work, Khayrallah and Koehn (2018) investigated the impact of different types of noise on the performance of machine translation systems. In the same year the WMT introduced a shared task on parallel corpus filtering, which went on to be repeated in the following years, and resulted in many different filtering approaches being proposed. In this work we aim to combine the recent achievements in data filtering with the original analysis of Khayrallah and Koehn (2018) and investigate whether state-of-the-art filtering systems are capable of removing all the suggested noise types. We observe that most of these types of noise can be detected with an accuracy of over 90% by modern filtering systems when operating in a well studied high resource setting. However, we also find that when confronted with more refined noise categories or when working with a less common language pair, the performance of the filtering systems is far from optimal, showing that there is still room for improvement in this area of research.

1 Introduction

The phenomenon of noisy data in the training of machine translation (MT) systems has been studied from various angles over recent years. To outline the impact of noise, Khayrallah and Koehn (2018) specified ten common noise categories and synthetically generated noisy data samples for each of them. By adding the noisy samples to an otherwise clean training corpus they measured the effect on the resulting translation system. Their conclusion was that neural machine translation (NMT) is less robust towards noisy data than statistical machine translation and that some noise types can prove very detrimental to NMT performance. As NMT had surpassed the statistical approaches just a few years prior, this work paved the way for a spiked interest in data filtering research for machine translation.

In the same year, the Conference on Machine Translation (WMT) started to host an annual shared task on parallel corpus filtering (Koehn et al., 2018, 2019, 2020) featuring a broad mix of academic and industrial submissions. This shared task highlights the general need for well working data filtering systems and resulted in the publication of a variety of new filtering approaches (Junczys-Dowmunt, 2018; Chaudhary et al., 2019; Lu et al., 2020). The WMT evaluations simulate a real-world data filtering task on web crawled data. Each participating system is required to extract a fixed amount of parallel data and is ranked according to the performance of the translation system that was trained on the selected data. While this form of evaluation is very relevant from a practical point of view, it does not show how well a certain approach is performing on detecting specific types of noise - information that is very important when working on improving data filtering approaches.

In this work we aim to unite both viewpoints in regards to data filtering. From the work of Khayrallah and Koehn (2018) we already know how detrimental certain categories of noise are to an NMT system, so we ask the question: *How well can state-of-the-art filtering systems distinguish the synthetic noise classes proposed by Khayrallah and Koehn (2018) from clean data?*

While downstream performance might be the ultimate objective for data filtering systems, this setup allows us to investigate the strengths and weaknesses of current data filtering systems. To further investigate the challenges for a filtering system we introduce several, more refined, synthetic noise categories and use them to benchmark the performance of the aforementioned filtering systems.

Findings of the Association for Computational Linguistics: ACL 2022, pages 2542 - 2551
May 22-27, 2022

2 Related Work

The task of data filtering for machine translation has attracted increasing attention in recent years, mainly for two reasons:

1. NMT models replaced the phrase based systems and it was shown that NMT models are less robust to many types of noise (Khayrallah and Koehn, 2018).

2. More and more parallel data is generated by web crawling techniques (Esplà-Gomis et al., 2019; Schwenk et al., 2019; El-Kishky et al., 2020; Schwenk et al., 2021) and this data is often quite 'noisy', making data filtering a crucial part of training competitive NMT systems.

To get an overview of existing approaches to data filtering for machine translation, a good place to start is the WMT shared task for parallel corpus filtering, which was held in 2018 (Koehn et al., 2018), 2019 (Koehn et al., 2019) and 2020 (Koehn et al., 2020). In these tasks, the participants are asked to select a fixed amount of data from a noisy parallel corpus using automatic methods. The examined language pairs were German-English (2018), Nepali-English (2019), Sinhala-English (2019), Khmer-English (2020) and Pashto-English (2020). The winning systems used a combination of language identification and language model and translation model scoring (Junczys-Dowmunt, 2018; Rossenbach et al., 2018), similarities in the cross-lingual sentence embedding space (Chaudhary et al., 2019) and even GPT-2 models (Lu et al., 2020). The only way in which data filtering systems are typically evaluated is by training a machine translation system on the selected data (Koehn et al., 2018, 2019, 2020). While this may be an intuitive evaluation criterion, it does not give many insights into the system performance regarding the detection of specific types of noise.

While much effort has been put into the building of powerful filtering systems, the same can not be said for analyzing their performance on specific types of noise. Belinkov and Bisk (2018) and Khayrallah and Koehn (2018) both examine the impact of various noise types on the performance of NMT systems but do not ask the question which types of noise can be handled reliably by data filtering systems. Xu and Koehn (2017) create synthetic noisy data to train a data filtering system with a classifier while Michel and Neubig (2018) create a 'noise translation benchmark' for NMT systems.

In this work we aim to fill this gap and systematically compare the performance of data filtering systems on specific categories of noise. We start from the categories defined in Khayrallah and Koehn (2018) and expand them further.

3 Types of Noise

We aim to investigate which noise categories can be reliably detected by state-of-the-art data filtering systems. Manual annotation of noisy corpora is expensive and tends to be very corpus and language specific, depending on the original data sources and the extraction techniques. Therefore we decide to investigate filtering systems on an array of noisy datasets mostly created synthetically from clean parallel data like it was done by Khayrallah and Koehn (2018). Most of the categories were introduced in the work of Khayrallah and Koehn (2018) (in the following marked with an asterisk (*)) but we propose two additional categories which we found to commonly occur in practice. Here we list all types of noise that we are investigating as well as our automatic and language agnostic methods of creating such noisy data. If a certain type of noise is specific to the source or target side of the data, we note the noisy side in brackets.

Misaligned Sentences* are created by shuffling the target side of a clean corpus. Hence, every source sentence get assigned a random target sentence from the same domain but (most probably) without any overlap in meaning.

Misordered Words (src|trg)* are obtained by arranging the words of either source or target sentence in a random order.

Wrong Language (src|trg)* samples are selected from a parallel corpus from a different language pair. We specify which side of the data is not fitting the intended task.

Untranslated (src|trg)* sentence pairs are created by converting a src-trg corpus into a src-src respectively trg-trg corpus via copying.

Short Segments (max. length)* are from a corpus with very short average sentence length. A segment being short does not imply that it is noisy or hurtful to the training. However we keep these categories for completeness sake in our analysis but do not emphasize on them in the experimental results.

Raw Crawled Data* is a mix of different types of noise and probably the most realistic noise category. We use data from an unfiltered web crawling corpus. Note that some sentence pairs from this category might be valid in practice and we address this in Section 5.1.

Over-/Undertranslation often times happen as a result of poor sentence splitting and alignment. To create sentence pairs in this category we remove the second half of the source sentence respectively target sentence.

Synthetic Translations can be found on an increasing number of websites and the simpler structure of synthetic translations (Edunov et al., 2018; Kim et al., 2017) can make them easier to be extracted and aligned by crawling scripts. To analyze if the the quality of the synthetic translations has an effect on detection accuracy, we extract human annotated data from the WMT shared task on news translation and group sentence pairs according to the human evaluation score.

It should be noted that most categories either distort the source or the target sentence of a pair. We keep this separation for our analysis even though many data filtering systems do not distinguish the languages direction, i.e. an X→Y filtering system can be used to clean a Y→X corpus. Hence, differences between the src- and trg-version of a noise category should best be seen as an indication of experimental variance or a dependency on the resources available for the two sides of a language pair.

4 Effect of Noise on NMT Performance

In this work we solely focus on the question how well certain types of noise can be detected by modern data filtering systems. However, an equally important question is, how detrimental a certain noise type is to the performance of an NMT system. In this section we briefly recapitulate the findings of Khayrallah and Koehn (2018), ranging the different types of noise according to their effect on NMT performance.

According to Khayrallah and Koehn (2018), the most severe noise type is the **Untranslated (trg)** category. Mixing just 20% of this type of noisy data into our clean training data results in an NMT performance drop to less than 10% BLEU, completely destroying translation performance. The authors explain this with the system learning to copy sentences rather than translating them into the correct language. The second worst type of noise is **Raw Crawled Data**, followed by **Misaligned Sentences**, **Misordered Words (src|trg)** and **Wrong Language (trg)** all leading to a significant performance degradation of the NMT system. On the other hand, Khayrallah and Koehn (2018) find that adding **Wrong Language (src)**, **Untranslated (src)** and **Short Segments (max. length)** leads to only minor performance degradation.

We additionally examine two types of noise which were not present in the work of Khayrallah and Koehn (2018), namely **Synthetic Translations** and **Over-/Undertranslation**. The latter mostly occurs as a result of bad segmentation and/or sentence splitting. We argue that in the most extreme case, this type of noise would coincide with the **Misaligned Sentences** so the impact of this category can be seen as an upper bound. In principle **Synthetic Translations** can be beneficial for NMT performance (Sennrich et al., 2016; Edunov et al., 2018; Kim et al., 2019). However, this depends heavily on the quality of the system used to generate this synthetic data and we argue that the purpose of web crawling is not to extract synthetic translation from possibly older machine translation models. Hence, we typically want to remove such samples. Therefore in this work we examine both the ability of the filtering systems to differentiate between good and bad synthetic data as well as to differentiate between synthetic and real parallel data.

We also point out that removing noise from the training data will have other benefits, aside from improved performance, such as faster convergence and less storage needs.

5 Experiments

5.1 Experimental Setup

In the following we briefly describe the filtering systems used in this study and the data conditions.

Filtering Systems

For our analysis we consider two of the strongest data filtering approaches to this date, based on either

- cross-entropy (CE) scores using translation and language models (Rossenbach et al., 2018).
- Language-Agnostic SEntence Representations (LASER) scores based on cross lin-

gual sentence embeddings (Chaudhary et al., 2019).

Both systems were among the winners of the WMT task on parallel corpus filtering in 2018 and 2019 respectively.

For the cross entropy system, we follow (Rossenbach et al., 2018) and train a source-to-target translation model $p_{s\rightarrow t}(e_1^I|f_1^J)$, a target-to-source translation model $p_{t\rightarrow s}(f_1^J|e_1^I)$ as well as two language models $p_s(f_1^J)$ and $p_t(e_1^I)$. We calculate a score for each sentence pair (f_1^J, e_1^I) with

$$\frac{1}{4}\Bigg(\frac{1}{J}\log p_s(f_1^J) + \frac{1}{I}\log p_t(e_1^I) \\ + \frac{1}{J}\log p_{t\rightarrow s}(f_1^J|e_1^I) + \frac{1}{I}\log p_{s\rightarrow t}(e_1^I|f_1^J)\Bigg).$$

The language models and translation models are implemented using the RETURNN toolkit (Zeyer et al., 2018). We use a 12 layer transformer model for the language models and the base transformer model (Vaswani et al., 2017) with 6 encoder and 6 decoder layers for the translation models.

For calculating the LASER scores, we generate cross-lingual sentence embeddings using the pretrained model provided by Artetxe and Schwenk (2019). The underlying system is trained as a multilingual translation system with a multi-layer bidirectional LSTM encoder and an LSTM decoder. No additional information about the input language is given to the encoder. The output vectors of the encoder are compressed into a single embedding of fixed length using max-pooling. This is the cross-lingual sentence embedding that the LASER model is generating. This vector is the only information about the input sentence which is transferred to the decoder. The intuition is, that two sentences with the same meaning but from different languages will be mapped onto the same embedding vector, as the translations that the decoder must produce should be identical. Once the cross-lingual sentence embeddings for every source and target sentence are extracted, we calculate the LASER scores for each sentence-pair according to Chaudhary et al. (2019):

$$\frac{2k\ \cos(f,e)}{\sum_{\hat{e}\in \mathrm{NN}_k(f)}\cos(f,\hat{e}) + \sum_{\hat{f}\in \mathrm{NN}_k(e)}\cos(\hat{f},e)}$$

where f, e are the sentence embeddings for source and target sentence respectively, $\cos(\bullet)$ is the cosine distance and $\mathrm{NN}_k(\bullet)$ is the set of the k nearest embeddings from the other language. The higher the LASER score of a sentence pair, the more similar the source and target sentence are semantically, corresponding to a better quality data point for training an NMT system.

Task	Data Type	#tokens (trg)	#lines
De→En	parallel	79M	3.1M
Km→En	parallel	4.9M	270k
	Km mono	546M	11M
	En mono	419M	11M

Table 1: Data resources used for the De→En and Km→En tasks.

Many filtering systems rely on some language identification (langID) toolkit as part of the selection method. With langID, each data point gets either a score of 1.0 or 0.0, depending on whether the predicted source and target languages match the given task or not. Since langID can be seen as a 'baseline' filtering technique on its own, we address it as a separate step in the filtering pipeline and denote for each experiment whether langID scores are included or not. When combining langID with the other methods, the scores are simply multiplied. For the task of language identification we use the popular `langid.py` toolkit (Lui and Baldwin, 2012).

Data

We benchmark the performance of the filtering systems on two language pairs: German→English and Khmer→English. For an overview over the amounts of data used we refer to Table 1. For both settings we remove a section of the parallel corpus from the training data to create the synthetic noise for the following categories: Misaligned Sentences, Misordered Words, Untranslated, and Over/Under-Translation.

Following Khayrallah and Koehn (2018), the De→En data consists of Europarl, News Commentary, and the Rapid EU Press Release corpus from the WMT2017 news translation task [1]. From this clean data we select randomly 7 × 50k sentence pairs to create the aforementioned synthetic noise categories. Every sentence pair contained in a noise category is excluded from the training corpus for all components of the data filtering systems. To create the remaining noise categories we used the same corpora as Khayrallah and Koehn (2018).

[1] `http://www.statmt.org/wmt17/translation-task.html`

Noise Category	Corrupted	Filtering Accuracy				
	Side	Cross	LASER	Language ID Filtering		
		Entropy		+ none	+ CE	+ LASER
Misaligned Sentences	none	65% / 65%	72% / 76%	50%	64% / 65%	71% / 75%
Misordered Words	src	89% / 89%	62% / 70%	50%	88% / 88%	61% / 70%
	tgt	95% / 96%	62% / 70%	50%	93% / 94%	61% / 70%
Wrong Language	src	89% / 89%	51% / 54%	97%	97% / 97%	97% / 97%
	trg	87% / 87%	54% / 60%	96%	96% / 96%	96% / 96%
Untranslated	src	62% / 62%	15% / 50%	97%	97% / 97%	97% / 97%
	trg	93% / 93%	14% / 50%	97%	97% / 97%	97% / 97%
Short Segments (≤ 2)	none	61% / 66%	62% / 69%	81%	83% / 85%	76% / 81%
Short Segments (≤ 5)	none	65% / 67%	59% / 64%	67%	73% / 75%	65% / 68%
Raw Crawl Data		94% / 95%	60% / 63%	84%	93% / 94%	79% / 84%
Overtranslation	src	67% / 67%	62% / 68%	52%	66% / 66%	62% / 68%
Undertranslation	trg	69% / 70%	64% / 70%	50%	68% / 68%	63% / 70%

Table 2: De→En Task: Accuracy of filtering methods when distinguishing different synthetic noise categories from clean, parallel data. Accuracies are reported a) in black: with knowledge of correct ratio between noisy and clean data b) in gray (oracle): with optimal noise-clean separation given the ranking of the filtering system.

For Km→En we use data from the WMT2020 parallel corpus filtering task[2]. We extract 20k sentence pairs from the clean corpus to create the synthetic noisy datasets. Since the Km→En task does not provide a lot of data to begin with, we use the same 20k sentence pairs to create the all synthetic noise categories and train the translation models for the cross-entropy filtering system on the remaining data. Since we were not able to find suitable data for the short segments and wrong language categories, we drop these for this language pair. Given that the parallel corpus is relatively small and of questionable quality, we additionally include all of the available monolingual Khmer data and subsample 11M English sentences to train the language models. To obtain raw crawled data we sample 20k sentence pairs from a web crawled corpus from the ParaCrawl project [3]. Note that this corpus also contains valid sentence pairs, however by manually annotating 150 sentence pairs, we observed that less than 10% of the sentence pairs were of acceptable quality.

5.2 Experimental Results

For each noise category described in Section 3, we generate a corresponding noisy testset. Each noisy testset is separately mixed together with an equal number of sentence-pairs sampled from a holdout set of the clean training data to create a mixed dataset where each sentence pair is labelled as either clean or noisy. To analyze the noise detection capability of the filtering systems we use either the cross-entropy or LASER approach to score each sentence-pair and sort the lines by score. For each system, a threshold-score is determined and all pairs with a score worse than the threshold are classified as noisy and all pairs with a score better than the threshold are classified as clean. This allows us to calculate the classification accuracy of the corresponding filtering system.

Two different thresholds are calculated separately for each mixed dataset:

- **Correct Ratio:** the threshold is chosen according to the true ratio between clean and noisy sentence pairs. If not denoted differently we use a 1:1 ratio.
- **Optimal (oracle):** the threshold is chosen such that we get the highest accuracy possible given the current filtering scores for the dataset. This requires knowledge of the true class and yields an upper bound of the filtering capabilities of the scoring system.

This means that if we were to rank the sentence pairs randomly, we would end up with around 50%

[2] http://www.statmt.org/wmt20/parallel-corpus-filtering.html

[3] https://paracrawl.eu/v7-1

Noise Category	Corrupted	Filtering Accuracy				
	Side	Cross	LASER	Language ID Filtering		
		Entropy		+none	+ CE	+ LASER
Misaligned Sentences	none	71% / 71%	72% / 72%	50%	62% / 65%	61% / 66%
Misordered Words	src	63% / 64%	53% / 54%	50%	57% / 62%	51% / 53%
	tgt	84% / 84%	50% / 51%	50%	69% / 76%	51% / 51%
Untranslated	src	69% / 70%	4% / 50%	86%	86% / 86%	86% / 86%
	trg	93% / 93%	2% / 50%	86%	86% / 86%	86% / 86%
Raw Crawl Data		77% / 77%	40% / 50%	71%	71% / 77%	70% / 71%
Overtranslation	src	56% / 56%	54% / 55%	51%	53% / 55%	52% / 54%
Undertranslation	trg	63% / 63%	61% / 61%	50%	58% / 60%	56% / 59%

Table 3: Km→En Task: Accuracy of filtering methods when distinguishing different synthetic noise categories from clean, parallel data. Accuracies are reported a) in black: with knowledge of correct ratio between noisy and clean data b) in gray (oracle): with optimal noise-clean separation given the ranking of the filtering system.

classification accuracy for both thresholds. Therefore 50% accuracy can be seen as a lower bound and if a filtering system drops significantly below that, its score is negatively correlated with data quality. When just using langID, each data point is scored with either 1.0 or 0.0, so no additional threshold calculation is needed. We note that we only focus on the quality of the data filtering scores and not and the question of how to select data, i.e. how to select a threshold. It is possible that this yields a more optimistic estimation of the filtering capabilities than is achievable in practice where even the clean to noise ratio is typically unknown.

German→English

The resulting accuracy scores for the De→En setup are listed in Table 2 where the **correct ratio** score is written on the left in each cell and the **optimal** score is written on the right and grayed out.

We find that the four noise categories that involve not-fitting languages, namely 'wrong language (src, trg)' and 'untranslated (src, trg)', are almost perfectly removed thanks to the language identification. Furthermore, langID also detects most of the noise stemming from raw crawl data and very short segments. As expected, langID fails to detect any noise coming from misalignment and over-/undertranslation. The cross-entropy approach has little trouble identifying the 'short segments ($\leq$ 2)', 'misordered words (src, trg)' and 'raw crawl data' noise types. The standalone LASER system fails to detect any noise stemming from incorrect languages which is compensated for by language identification filtering. From the noise categories that are defined in Khayrallah and Koehn (2018), only 'misaligned sentences' and 'short segments ($\leq$ 5)' pose a serious detection problem. While for the latter, one could argue that it is neither harmful to the system performance (Khayrallah and Koehn, 2018) nor actually noise, the bad performance on the 'misaligned sentences' is quite surprising as this type of noise is quite severe and should be detected quite reliably in theory by both cross entropy and LASER filtering. The noise categories 'overtranslation' and 'undertranslation', which are newly added in this work, pose a serious problem for all filtering methods. In general, there is not much difference in accuracy for the combined filtering systems when detecting sentence corruptions on the German side (source) compared to the English side (target).

Regarding the selection method, we find that the cross-entropy approach is less susceptible towards different types of thresholds compared to the LASER approach. It is not clear to us why this is the case but we speculate it has to do with the fact that the CE approach is less reliant on the langID scores overall. Both cross-entropy and LASER benefit from the combination with langID.

Khmer→English

As the second language pair, we use Khmer→English, a task where the languages have very little in common in terms of syntax and data resources are scarce. The accuracy scores for the Km→En setup are listed in Table 3. As one might expect, the language identification does not perform as good for the Khmer language.

In fact in some cases the inclusion of langID is actively hurting the overall filtering performance, for example for the ‘misaligned sentences’ and ‘undertranslation’ categories. In contrast to the De→En setting, most noise categories can not be detected reliably, with the exception of the ‘untranslated (src, tgt)’, ‘misordered words (tgt)’ and ‘raw crawl data’. Most of the times, noise on the English target side can be detected more reliably than on the Khmer source side, although still not with a very good accuracy.

Synthetic Data Detection

Next, we investigate another type of noise that is often overlooked, namely synthetic data where either the source or the target side is created by MT systems. To obtain the synthetic data as well as corresponding quality annotation, we use the human-scored automatic translations of the WMT De→En news translation task from 2016 to 2019. For each year, we rank all hypotheses according to the score of the human annotators. We take the the worst 30% of translations as our noisy data. As clean data, on the one hand we take the best 30% of translations and on the other hand we take the reference translations generated by professional human translators.

In Table 4 the resulting filtering accuracy is shown for differentiating between the 30 % best scored and the 30 % worst scores translations as well as between the 30 % worst scored translations and the (human) reference translations. Interestingly the systems have a harder time differentiating between good translations of Mt systems and humans compared to differentiating between good and bad automatic translations. However, we find that the filtering systems can not reliably differentiate between synthetic and human translations, nor between a good and a bad synthetic translation from the same domain.

Adversarial Data	Filtering Accuracy	
	Cross Entropy	LASER
synthetic & high quality	62% / 62%	62% / 63%
references	54% / 55%	44% / 50%

Table 4: Filtering accuracy of two data filtering systems. Systems are required to distinguish synthetic translation, with poor human-rating from adversarial data. ‘synthetic & high quality’ comes from the same test set (but obtained best human scores), ‘references’ are the official references (from humans) of the same test set.

Mixed-Noise Categories

Apart from analysing the performance of the filtering systems on the individual noise categories, we briefly look at the performance on a combined dataset which consists of the concatenation of all individual noisy datasets (equal ratio of clean and noisy data).

The results for both language pairs are shown in Table 5. We again see that the systems perform better on De→En compared to Km→En which is mainly due to langID performing significantly worse on Km→En. In fact, the cross-entropy approach performs batter as a standalone system rather than in combination with langID.

Lastly we test the filtering capabilities in an extremely noisy scenario and report the results in Table 6. Note that this data set exhibits a #clean:#noise ratio of 1:12 for De→En and 1:8 for Km→En. Since this ratio is used in the filtering system (to set a filtering threshold for the ‘correct ratio’) the filtering accuracy system will always correctly classify at least a fraction of

$$\frac{\#\text{noise} - \#\text{clean}}{\#\text{noise} + \#\text{clean}}.$$

Analyzing the performance of the filtering systems using an ‘optimal’ threshold value (gray values in Table 6) we noticed that they classify all sentence pairs as noise. Since the data distribution is very biased towards noisy data we also report F1-scores for this experiments in Table 7. We find that both systems are doing a poor job at noise detection if the clean-to-noisy data ratio gets too small.

6 Conclusion

The aim of this work is to determine how well state-of-the-art data filtering systems can detect different types of noise common in parallel machine translation datasets. We create synthetic noisy datasets for all noise categories defined by Khayrallah and Koehn (2018) as well as for additional noise types that we define in this work. We find that modern data filtering systems can detect most types of noise with an accuracy of well over 90% on a German→English task, that features a medium sized, rather clean training corpus for the filtering systems.

Language Pair	Filtering Accuracy				
	Cross Entropy	LASER	Language ID Filtering		
			+none	+ CE	+ LASER
De→En	75% / 76%	49% / 50%	73%	81% / 82%	68% / 73%
Km→En	70% / 70%	46% / 51%	61%	67% / 68%	61% / 62%

Table 5: Filtering accuracy on two language pairs with a clean-noise ratio of 1:1 for De→En and Km→En by limiting the size of each noise category before ensembling all noise categories from Table 2 respectively Table 3.

Language Pair	Minimal Accuracy	Filtering Accuracy				
		Cross Entropy	LASER	Language ID Filtering		
				+none	+ CE	+ LASER
De→En	85%	89% / 92%	85% / 92%	55%	91% / 92%	87% / 92%
Km→En	78%	85% / 89%	79% / 89%	54%	86% / 89%	83% / 89%

Table 6: Filtering accuracy on two language pairs with high a clean-noise ratio of 1:12 for De→En and 1:8 for Km→En by combining all noise categories from Table 2 respectively Table 3. Note, that if these very biased distributions are accessible to the filtering system, a minimal accuracy can be guaranteed (Column 2) except for the case of pure langID filtering (since it does not rely on the data ratio).

Language Pair	F1-Score				
	Cross Entropy	LASER	Language ID Filtering		
			+none	+ CE	+ LASER
De→En	94% / 96%	92% / 96%	68%	95% / 96%	93% / 96%
Km→En	92% / 94%	88% / 94%	67%	92% / 94%	90% / 94%

Table 7: Filtering performance based on F1-score on two language pairs with high a clean-noise ratio of 1:12 for De→En and 1:8 for Km→En by combining all noise categories from Table 2 respectively Table 3.

However, well-formed but misaligned sentence pairs and over-/undertranslation can only be detected with an accuracy of less than 70%. When it comes to detecting more subtle errors like distinguishing between a good and a poor synthetic translation, the systems exhibit even worse performance. Furthermore, when switching to a less common language pair, namely Khmer→English, the performance of the filtering systems degrades significantly compared to German→English. In conclusion we find that the task of data filtering as defined by Khayrallah and Koehn (2018) is not yet solved. There is still much room for improvement, especially when going to more subtle types of noise or to less common language pairs.

For future research or when applying data filtering for a downstream task, we want to emphasize the following points:

- For high resource languages, langID is a good basis to start from. Subsequent filtering steps should specifically focus on phenomena that langID can not detect such as misaligned sentences and over-/undertranslation.

- For low resource languages, it might be beneficial to drop the langID filtering step, if the subsequent methods have their own (implicit) ways of detecting wrong language. It might be helpful to train some language classifier yourself if in-domain monolingual training data is available.

- Even when (roughly) knowing the percentage of noise in the data, removing this percentage is most of the times not the optimal choice in terms of filtering accuracy. Alternative methods such as a fixed score threshold independent of the selected percentage should also be considered.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 694537, project "SEQCLAS"). The work reflects only the authors' views and the European Research Council Executive Agency (ERCEA) is not responsible for any use that may be made of the information it contains.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, 7:597–610.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

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, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 261–266.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Miquel Esplà-Gomis, Mikel L 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 Volume 2: Translator, Project and User Tracks*, pages 118–119.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 888–895.

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, NMT@ACL 2018, Melbourne, Australia, July 20, 2018*, pages 74–83. Association for Computational Linguistics.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Yunsu Kim, Julian Schamper, and Hermann Ney. 2017. Unsupervised training for large vocabulary translation using sparse lexicon and word classes. In *Conference of the European Chapter of the Association for Computational Linguistics*, pages 650–656, Valencia, Spain. [poster].

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT2020 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, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 3: Shared Task Papers, Day 2*, pages 54–72.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 739–752, Belgium, Brussels. 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.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the System Demonstrations, July 10, 2012, Jeju Island, Korea*, pages 25–30.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553.

Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney. 2018. The RWTH Aachen University filtering system for the WMT 2018 parallel corpus filtering task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 946–954, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950.

Albert Zeyer, Tamer Alkhouli, and Hermann Ney. 2018. RETURNN as a generic flexible neural toolkit with application to translation and speech recognition. In *Annual Meeting of the Assoc. for Computational Linguistics*, volume abs/1805.05225.