Bifixer and Bicleaner: two open-source tools to clean your parallel data

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, Sergio Ortiz Rojas
Prompsit Language Engineering, S.L.
Campus UMH, Edifici Quorum III
Av. de la Universitat, s/n. 03202. Elx. Spain
{gramirez,jzaragoza,mbanon,sortiz}@prompsit.com

Abstract

This paper shows the utility of two open-source tools designed for parallel data cleaning: Bifixer and Bicleaner. Already used to clean highly noisy parallel content from crawled multilingual websites, we evaluate their performance in a different scenario: cleaning publicly available corpora commonly used to train machine translation systems. We choose four English–Portuguese corpora which we plan to use internally to compute paraphrases at a later stage. We clean the four corpora using both tools, which are described in detail, and analyse the effect of some of the cleaning steps on them. We then compare machine translation training times and quality before and after cleaning these corpora, showing a positive impact particularly for the noisiest ones.

1 Introduction

Parallel corpora are usually the main source of information used to learn machine translation models. The availability of corpora has encouraged the advance of machine translation in both academy and industry settings. Publicly available parallel corpora (Europarl, News Commentary, United Nations, etc.) have been used for decades now, not only to produce machine translation but also other by-products such as dictionaries, concordances, synonyms, paraphrases, etc. In machine translation, due to the ability of statistical models to hide imperfections without statistical significance, filtering out noise from these corpora was not very important. Now that neural models have superseded statistical ones, we need to be more careful about noise in the input as it has a higher impact on the output, as discussed in (Khayrallah and Koehn, 2018) and (Rikters, 2018).

Inspired by recent work on filtering parallel corpora to maximize the quality of machine translation from the shared tasks organised at WMT18[1] and WMT19[2], we review how noisy some of the most popular or recent publicly available corpora are and how this impacts the quality of the output of state-of-the-art neural machine translation. Our motivation is twofold: getting high-quality monolingual and bilingual data and getting high-quality machine translation for English–Portuguese. We will further use this resources to compute paraphrases in the framework of a research project.

In order to inspect and filter out noise, we use Bifixer and Bicleaner,[3] a couple of publicly available cleaning tools released as part of the ParaCrawl European project.[4] These tools have been mainly used to filter out noise from the raw version of automatically crawled parallel corpora in more than 30 language combinations. Here we use them in a very different scenario: we take already released publicly available corpora, either widely used in the past or recent. We analyse the main problems of the corpora and review the cleaning steps and their impact on the final size of the corpora.

[1] http://www.statmt.org/wmt18/parallel-corpus-filtering.html
[2] http://www.statmt.org/wmt19/parallel-corpus-filtering.html
[3] Code available at https://github.com/bitextor/bifixer and https://github.com/bitextor/bicleaner
[4] See more info and available corpora on https://paracrawl.eu

To evaluate the effect of cleaning, we train neural machine translation systems before and after filtering them and report both performance results and evaluation through automatic metrics. We do so for English and Portuguese in both translation directions as Portuguese is one of the target languages in our research project related to paraphrasing. Our focus for this paper is, though, the evaluation of the cleaning tools intrinsically and extrinsically through machine translation.

The rest of the paper is organised as follows: in section 2 we discuss the cleaning steps applied to the corpora and analyse the type of noise found in them; section 3 describes the MT experiments and reports on the results; finally, section 4 depicts the conclusions and some ideas for future work.

2 Cleaning parallel corpora

Although parallel corpora cleaning has been explored in previous works, the most recent state-of-the-art can be found as part of the findings of the shared tasks on Parallel Corpora Filtering in WMT18 (Koehn et al., 2018) and WMT19 (Koehn et al., 2019). Participants in these shared tasks applied a bunch of techniques looking for high-quality data inside noisy corpora. Most of these techniques are a mixture of pre-filtering rules for obvious noise, scoring functions of all sorts (language models, neural translation models, etc.) and classification to discriminate between high-quality and low-quality sentence pairs. Diverse techniques have been applied to both high-resource and low-resource languages.

The results encourage filtering, especially for high-resource scenarios involving neural machine translation. On the other hand, no clear trend was devised for low-resource scenarios nor for statistical machine translation.

Some of these techniques have been already implemented and evaluated in Bicleaner (Sánchez-Cartagena et al., 2018). Bifixer adds a different way of exploring corpora cleaning: restorative cleaning. With this step, we aim at fixing content and getting unique parallel sentences before filtering out noise.

2.1 Cleaning by restoring

The first step taken for corpora cleaning in the most recent ParaCrawl pipeline is restorative cleaning. It is performed by Bifixer. Currently, the following sub-steps are applied to the sentences of an input parallel corpus:

- **empty side removal**: lines without content in either source or target are removed
- **character fixing**: sentences with encoding issues (Mojibake), HTML entities issues, wrong alphabet characters and space or punctuation issues are fixed
- **orthography fixing**: words with frequent and straightforward typos are rewritten. It is currently available for Danish, German, English, Spanish, Dutch, Norwegian, Portuguese and Turkish
- **re-splitting**: using NLTK[5] on sentences over 15 tokens by default, and taking into account source and target, re-splitting is applied. Only if the number of splits is equal on both sides, the new splitting is kept, otherwise the original one remains.
- **duplicates identification**: a hash identifier is calculated and added to each pair of sentences in order to identify both duplicate and, optionally, near-duplicate (i.e. ignoring casing, accents, diacritics and digits) parallel sentences. A score is calculated in order to decide the best near-duplicate to be chosen. We will apply both duplicate and near-duplicate marking in our experiments. [6]

The rationale behind the steps performed by Bifixer is to have the best possible content for machine translation: fixing encoding or typos will produce a more consistent content; too long sentences by themselves or because they are two glued sentences, are normally discarded from training sets; finally, duplicates and near duplicates are poor content to be given to learning systems.

2.2 Cleaning by filtering

After restorative cleaning, sentence pairs are sent to Bicleaner, a parallel sentence noise filter and classifier tool. Bicleaner was first released in 2018 as part of the ParaCrawl software, and has been used outside the project in several works such as (Morishita et al., 2019), (Defauw et al., 2019) and (Chaudhary et al., 2019). The tool performs the following sub-steps:

[5] `https://www.nltk.org/`

[6] Please note that Bifixer will not actually remove the duplicates, it will just mark them. An additional processing needs to be added for removal.

1. **Pre-filtering** based on rules is the first step in Bicleaner. There is a set of 37 rules currently implemented. Some of the rules are language-dependent and use language identification based on CLD2[7] for filtering. While some of them look into one of the sides of the corpus, some others take into account both sides. In general, they filter obvious noise such as sentences with a very different length in source and target. They were designed to target noise from web crawled content but most of the rules apply to any corpus. We do perform an analysis of the most productive rules for different scenarios in subsection 2.3. When a pair of sentences matches a rule in this step, it is set as "0" score, meaning that it should be discarded.

2. **Language model fluency scoring** allows filtering in a more refined way. It is language-dependent and uses a character-based language model. Using characters instead of n-grams reduces the amount of data needed to train the model, although it limits the usage for some languages with very scarce resources or special alphabets. The fluency filter provides a score for each sentence pair against the language model. Only pairs below a set threshold (0.5) are matched to a 0 score for the rest of the pipeline, meaning that they must be discarded. Recently, the fluency filter was integrated as the last pre-filtering rule in the workflow. This step will be disabled for our experiments in this paper as it is mainly intended for very big-sized corpora.

3. **Classification** based on a random-forest machine learning model is the last step. The classifier takes all sentences not marked with a score of 0 from previous steps and classifies them by providing a score between 0 (bad) and 1 (good). The official ParaCrawl released corpora contain only sentences above a score of 0.7. Other studies have reported better machine translation scores using sentences above 0.5. We will explore both thresholds in our experiments.

2.3 Applying cleaning to popular corpora

To better understand the effect of cleaning, we take four corpora from the bunch of publicly available parallel ones in English and Portuguese. Except for WikiMatrix[8], all of them are taken from OPUS:[9]

- Europarl, version 7, (Koehn, 2005): it is a widespread used corpus in machine translation, last released in 2011, containing parallel sentences from the proceedings of the European Parliament. This version for English–Portuguese contains 2.2 million sentences.

- OpenSubtitles 2018, version 6, (Lison and Tiedemann, 2016): this is also a very popular corpus. It comes mostly from volunteers translating subtitles on the net.[10] The last version from 2018 contains more than 33 million parallel sentences for English–Portuguese.

- JW300 (Agic and Vulic, 2019): it is a very recent corpus with only one version released. It was compiled by crawling the `jw.org` website and contains 2.1 million sentences in English–Portuguese.

- WikiMatrix (Schwenk et al., 2019): also recently released, it is an effort to compile translations found in Wikipedia. The corpus in English–Portuguese contains 4.4 million of parallel sentences.

In our setting, Bifixer was used without modifications applying deduplication also for near duplicates.[11]. Bicleaner provides pre-trained classifiers for many languages including English–Portuguese,[12]. But, in order to avoid misleading results, we trained new models leaving out the corpora that we intend to analyse.[13] Corpora, training times and sizes are compiled in Table 1. Training corpora are all taken from and cleaned with Bifixer and the pre-filtering rules step in Bicleaner before training. The training of Bicleaner models has been run in an Intel Core i9 using 32 cores and the cleaning of corpora has been run in an Intel Core i7 using 8 cores.

[7] `https://github.com/CLD2Owners/cld2`

[8] `https://github.com/facebookresearch/LASER/tree/master/tasks/WikiMatrix`

[9] `http://opus.nlpl.eu/`

[10] `http://www.opensubtitles.org/`

[11] Cloned from Github on 17th February 2020

[12] `https://github.com/bitextor/bicleaner-data/releases/tag/v1.3`

[13] To train our special models for Bicleaner, we follow the guidelines in `https://github.com/bitextor/bicleaner/wiki/How-to-train-your-Bicleaner`

After completing the training step, we apply Bifixer and Bicleaner to the selected corpora. Firstly, after Bifixer, we observe that:

- **we get more data**: by mean, 1,1% new sentences are recovered after re-splitting. The impact on size is almost negligible but it will be noticeable in the quality of the final subset of sentences retained.

- **we keep only unique data**: by mean, 9.8% duplicates or near-duplicates are removed. The biggest impact can be seen in OpenSubtitles (8.1 million sentences representing 24.5% of the whole corpus are removed). It is also noticeable in JW300 where more than 10% is removed at this step.

- **we get a better output**: by mean, 4.6% of the sentences have been fixed (typos, encoding, HTML entities, trailing spaces, etc.) as described in section 2.1.

Secondly, from Bicleaner pre-filtering rules[14] we observe that:

- **most of the content is still retained after pre-filtering rules**: by mean, 85.7% goes to the classifier step. It drops down to 67.7% for OpenSubtitles and is as high as 96.3% for Europarl.

- **none of the corpora matches all the 37 pre-filtering rules**: WikiMatrix matches 35, OpenSubtitles matches 33, Europarl 28 and JW300 only 25.

- **the main source of noise is equivalent across corpora**: it comes from sentences with language identification issues (both source and target are in the same language, the identified languages are not reliable), length issues (unusual length ratio between source and target, sentences have just 1 or 2 tokens in both sides) or quality issues (sentences contain mainly non-alphabetic characters).

It is worth comparing this analysis to the one obtained from ParaCrawl raw files. The raw files contain preliminary and very noisy candidate parallel sentences from crawled websites. For English–Portuguese, version 5 of the raw corpus,[15] a significantly smaller portion is retained after pre-filtering rules compared to our current scenario: only 27.2% of the raw corpus goes to the classifier step. The main reasons why sentences are removed are, though, very similar to the ones applying to the corpora in this paper. From all the 37 rules matched:

- 25.8% is removed by rules matching length issues: very short sentences (only 1 or 2 tokens on both sides) from web crawled content are often badly aligned and of poor quality. On the other side, very long ones (more than 1024 bytes) are often problematic. Too odd length ratios are the cause of the removal of 9,7% of the content.

- 19.5% is removed by rules matching language identification or encoding issues: same language on both sides, languages unreliably identified and characters out of the range of Unicode char sets

- 15.2% is removed by rules matching quality issues: sentences are mainly symbols or URLs, upper and lowercase distribution is odd on either side, match code-like patterns, contain poor language, etc.

- additionally, 9.3% is removed because source and target are identical or just differ in numbers and punctuation

For our four corpora, the last step is scoring with Bicleaner classifier. After classification, we filter sentence pairs below a couple of thresholds: 0.5 and 0.7. For the most aggressive threshold, 0.7, we remove by mean 22.9% of the corpora, being WikiMatrix the most impacted corpus by this step, with a 37.3% of discarded sentences, followed by OpenSubtitles (32.5%). With the threshold set to 0.5, the removal drops to a mean of 10.9%. In this scenario, WikiMatrix loses 21.1% of the corpus, followed by OpenSubtitles (15.5%).

In all (see table 2), for the most aggressive cleaning, the 0.7 classifier score scenario, we observe that the initial sizes of the corpora are reduced by a mean of 37.2% after applying Bifixer and Bicleaner. OpenSubtitles is the corpus with

[14] We disable the fluency filter for our experiments, as it is mainly intended for very big-sized corpora.

[15] www.paracrawl.eu/releases

the biggest percentage of removals which represents 64.8% of the total. The classifier step is the most frequent reason for discarding sentences. Europarl is the corpus with the smallest percentage of removals, only 12.5% of the corpus is lost during cleaning.

2.4 Quick evaluation of cleaned data

After cleaning, we sample 100 random sentences from each corpus and manually annotate them with KEOPS,[16] an open-source tool which provides a framework for manual evaluation of parallel sentences. KEOPS was also released as part of the ParaCrawl project. Error annotation is done following the European Language Resource Coordination (ELRC) validation guidelines. [17] We annotate each sentence pair as Valid or as containing one of the following 7 errors: Wrong Language Identification, Wrong Alignment, Wrong Tokenization, Machine Translation, Translation Error or Free translation.

From manual annotation, we get the following insight:

- in Europarl only 2 sentences out of 100 present issues with sentence splitting either in source or in target.
- in JW300 we discover an issue with the original tokenization: hyphens and quotes are separated from the words they belong to (e.g. `lembra - se "` instead of `lembra-se"` in Portuguese). Ignoring those, only 2 sentences are badly split, 2 contain translation errors and 3 are too free translations.
- in OpenSubtitles, 11 sentences out of 100 present issues: 5 are badly tokenized, 2 are clearly bad machine translations and 4 are too free translations.
- in WikiMatrix, 30 sentences out of 100 present issues: 7 are miss-aligned, 4 are badly tokenized, 5 contain bad machine translation, 10 contain translation errors and 4 are too free translations.

These results show room for improvement for the cleaning tools that will be taken into account as future work. They also give an idea of the characteristics of the corpus, a valuable piece of information to keep in mind when selecting corpora for a number of natural language processing tasks.

3 Evaluation through machine translation

In order to evaluate the impact of cleaning, we train neural machine translation systems before and after cleaning for each of the four corpora inspected. This allows us to see if better and reduced versions of the corpora produce a better machine translation output. We measure the impact of cleaning in the output by using automatic metrics. We also measure training times to see if size reduction and a more consistent content leads to a more efficient training process.

Machine translation systems are trained on each corpus before and after cleaning, for both translation directions and for both 0.7 and 0.5 Bicleaner thresholds. We train Transformer-base models with 32,000 vocabulary using Marian (Junczys-Dowmunt et al., 2018) and SentencePiece. We use development and test sets from TED Talks proposed by (Ye et al., 2018) and report BLEU scores computed with sacreBLEU[18]. Results for BLEU scores for all the 24 systems are reported in table 3 while training times are shown in table 4.

From the results, we can see that cleaning has a positive impact on all the corpora, both in speeding up training times and in slightly improving BLEU scores for almost all corpora and translation directions: only Europarl, English-Portuguese, just stays the same. Thus, no degradation is introduced with corpora size reduction, but rather the opposite: the most aggressive cleaning (0.7) scenario, leading to the smallest corpora sizes, gets consistently better BLEU scores for all the experiments. This scenario leads also to the best training times in most cases. Indeed, the highest improvements in BLEU scores (from +1 to +2.2 absolute BLEU points) are obtained when 22M sentences (two-thirds of the corpus) are filtered out from OpenSubtitles.

4 Conclusions

We have applied Bifixer and Bicleaner, two open-source tools built inside the ParaCrawl project, to

[16] `Downloadthecodefromhttps://github.com/paracrawl/keops`

[17] `http://www.lr-coordination.eu/`

[18] Signature: `BLEU+case.mixed+lang.pt-en+numrefs.1 +smooth.exp+tok.13a+version.1.4.2`

clean four publicly available parallel corpora for English–Portuguese. After a review of the tools and the cleaning steps performed to the four corpora, we evaluate the output of neural machine translation before and after cleaning them to see their impact.

Cleaning reduces the size of the corpora. For some of them (Europarl, JW300), the reduction is low but for others, cleaning removes half of the corpus (WikiMatrix) or up to two thirds (OpenSubtitles).

Cleaned corpora, in the most aggressive cleaning scenario (Bicleaner scores above 0.7), lead consistently to equal or slightly better results for BLEU scores in machine translation, not degrading the results in any case and speeding up machine translation training times.

At bigger scale (more languages, bigger sizes for all corpora together) all this could result in remarkable savings of disk space and training times without compromising machine translation quality and producing higher-quality corpora.

Both tools can be currently used without any further effort for more than 30 language combinations and prove to be a cheap and effective step before using parallel corpora for machine translation or other natural language processing tasks. For non-supported languages, Bifixer will only require a list of monolingual safe replacements for typos. Bicleaner, though, will require training resources and time, although much less than other methods.

From a closer look, we observe that, for less noisy corpora as Europarl, some of the Bicleaner pre-filtering rules are too severe and could probably be relaxed. In particular, the removal of too short sentences should be further inspected for already high-quality data.

As further work, although Bifixer and Bicleaner have been used for many other languages inside the ParaCrawl project, it would be interesting to validate the results obtained in this paper for other language combinations and corpora.

Outside machine translation, we believe that cleaning is also good for other tasks such as improving sentence alignment or paraphrase extraction. Both, and specially paraphrase extraction, will be explored as further work as part of a research project that will use the results of this paper as best practices to pre-process corpora.

Acknowledgment

Work supported by project ParaCrawl, actions number 2017-EU-IA-0178 and 2018-EU-IA-0063, funded under the Automated Translation CEF Telecom instrument managed by INEA at the European Commission. Also supported by the Spanish research program *Impulso a las Tecnologías habilitadoras digitales*, action number TS1-100905-2019-4 from the Secretary of State for Digitalisation and Artificial Intelligence currently under the Ministry of Economic Affairs and Digital Transformation. We thank Carmen Iniesta López for her valuable feedback and suggestions.

References

Agic, Zeljko and Ivan Vulic. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In Korhonen, Anna, David R. Traum, and Lluís Màrquez, editors, *ACL (1)*, pages 3204–3210. Association for Computational Linguistics.

Chaudhary, Vishrav, 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, August. Association for Computational Linguistics.

Defauw, Arne, Tom Vanallemeersch, Sara Szoc, Frederic Everaert, Koen Van Winckel, Kim Scholte, Joris Brabers, and Joachim Van den Bogaert. 2019. Collecting domain specific data for mt: an evaluation of the paracrawl pipeline. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 186–195.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

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

Koehn, Philipp, 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*, Belgium, Brussels, October. Association for Computational Linguistics.

Koehn, Philipp, 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. Association for Computational Linguistics.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Lison, Pierre and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *LREC*. European Language Resources Association (ELRA).

Morishita, Makoto, Jun Suzuki, and Masaaki Nagata. 2019. Jparacrawl: A large scale web-based english-japanese parallel corpus. *arXiv preprint arXiv:1911.10668*.

Rikters, Matīss. 2018. Impact of Corpora Quality on Neural Machine Translation. In *In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018)*, Tartu, Estonia.

Sánchez-Cartagena, Víctor M., Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium, October. Association for Computational Linguistics.

Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Ye, Qi, Sachan Devendra, Felix Matthieu, Padmanabhan Sarguna, and Neubig Graham. 2018. When and why are pre-trained word embeddings useful for neural machine translation. In *HLT-NAACL*.

Bicleaner	Corpora	Subset (#sentences)		Time
Training: probabilistic dictionaries	**DGT**	3.5M	10M	10h
	EuBookshop	3.5M		
	JRC	1M		
	Capes	1M		
	Tilde	1M		
Training: classifier model	**SCielo**	50k	100k	10’
	NewsCommentary	25k		
	Global Voices	25k		
Cleaning: pre-filtering and classifying	**Europarl**	2.2M		20’
	JW300	2.1M		20’
	WikiMatrix	4.4M		46’
	OpenSubtitles	33.2M		112’

Table 1: Corpora, sizes, training and cleaning times for Bicleaner.

			Europarl		JW300		WikiMatrix		OpenSubtitles	
			# sent	%	# sent	%	# sent	%	# sent	%
	Original		2,002,943	100	2,102,425	100	4,458,124	100	33,222,606	100
Bicleaner Bifixer	**Re-split**		+17,648	+0.8	+55,874	+2.6	+39,670	+0.8	+9,951	+0.03
	Dedup		-65,728	-3.2	-228,043	-10.8	-32,151	-0.7	-8,157,302	-24.5
	Pre-filter		-25,755	-1.2	-91,659	-4.3	-392,648	-8.8	-2,594,724	-7.8
	Classify	**0.5**	-40,814	-2.0	-105,385	-5.0	-941,887	-21.1	-5,151,005	-15.5
		0.7	-177,825	-8.8	-275,696	-13.1	-1,666,923	-37.3	-10,814,463	-32.5
	Total	**0.5**	-114,649	-5.7	-369,213	-17.5	-1,327,016	-29.7	-15,893,080	-47.8
		0.7	-251,660	-12.5	-539,524	-25.6	-2,052,052	-46.0	-21,556,538	-64.8
	Final	**0.5**	1,888,294	94.2	1,733,212	82.4	3,131,108	70.2	17,329,526	52.16
		0.7	1,751,283	87.4	1,562,901	74.3	2,406,072,00	53.9	11,666,068	35.1

Table 2: Number of sentences added (+) or removed (-) after each cleaning step.

		Europarl			JW300			WikiMatrix			OpenSubtitles		
		size	BLEU score		size	BLEU score		size	BLEU score		size	BLEU score	
			en–pt	pt–en		en–pt	pt–en		en–pt	pt–en		en–pt	pt–en
Before cleaning		*2.2*	**26.2**	31.5	*2.1*	29.0	34.1	*4.4*	35.8	36.8	*33.2*	31.2	37.9
After cleaning	**0.5**	*1.8*	26.0	31.5	*1.7*	29.1	34.2	*3.1*	36.2	36.8	*17.3*	31.9	39.5
	0.7	*1.7*	**26.2**	**31.7**	*1.5*	**29.4**	**34.4**	*2.4*	**36.3**	**37.0**	*11.6*	**32.2**	**40.1**

Table 3: BLEU scores for all NMT systems trained after and before cleaning in both translation directions and for two different Bicleaner classifier thresholds. Sizes of corpora are provided. Best NMT systems are shown in bold.

		Europarl			JW300			WikiMatrix			OpenSubtitles		
		size	training time		size	training time		size	training time		size	training time	
			en–pt	pt–en		en–pt	pt–en		en–pt	pt–en		en–pt	pt–en
Before cleaning		*2.2*	21.6	20.4	*2.1*	18.4	17.7	*4.4*	28.4	36.2	*33.2*	54.7	44.1
After cleaning	**0.5**	*1.8*	20.5	**18.7**	*1.7*	16.2	**18.4**	*3.1*	29.3	29.5	*17.3*	25.9	**33.3**
	0.7	*1.7*	**18.8**	21.6	*1.5*	**13.3**	**18.4**	*2.4*	**23.9**	**26.8**	*11.6*	**22.1**	33.4

Table 4: Training times in hours for all the NMT systems. Sizes of corpora are provided. Best training times are shown in bold.