Human evaluation of web-crawled parallel corpora for machine translation

Gema Ramírez-Sánchez, Marta Bañón
Jaume Zaragoza-Bernabeu, Sergio Ortiz-Rojas
Prompsit Language Engineering, Spain
{gramirez, mbanon, jzaragoza, sortiz}@prompsit.com

Abstract

Quality assessment has been an ongoing activity of the series of ParaCrawl efforts to crawl massive amounts of parallel data from multilingual websites for 29 languages. The goal of ParaCrawl is to get parallel data that is good for machine translation. To prove so, both, automatic (extrinsic) and human (intrinsic and extrinsic) evaluation tasks have been included as part of the quality assessment activity of the project. We sum up the various methods followed to address these evaluation tasks for the web-crawled corpora produced and their results. We review their advantages and disadvantages for the final goal of the ParaCrawl project and the related ongoing project MaCoCu.

1 Introduction

Machine translation and particularly neural machine translation is a data hungry process. Data, ideally in the form of parallel texts, is many times scarce for many languages, poorly varied for others or very low quality. Multilingual websites are a great source of parallel data to complement these poor data scenarios, enabling the use and usefulness of machine translation for many use cases. But the web is wild and automatic harvesting of parallel data is not exempt of errors.

Web-crawled parallel content, usually noisy, can be then filtered for quality. The final parallel sentences that make it to a web-crawled parallel corpus will have gone through a complex pipeline before they are compiled and released in the form of a parallel corpus.

Once produced, how good are these parallel sentences? How good is the corpus as a whole? What kind of errors does it contain? Are these errors problematic for building machine translation? What type of evaluation process can help us to identify action points to improve the production pipeline?

These are the questions that we were trying to answer when designing the tasks that would be carried out as part of the quality assessment activity in the ParaCrawl project. (Bañón et al., 2020) provides a full description of the project, methods to gather corpora and a description of released corpora and their usefulness to create machine translation systems. ParaCrawl goal was the release of the largest collection of parallel corpora harvested from multilingual websites to advance machine translation. Initially targeting 23 co-official European languages paired with English, the final version contains also Norwegian Nynorsk, Norwegian Bokmål and Icelandic paired with English and 3 corpora for co-official languages in Spain paired with Spanish. Version 9 accounts for 1.457 million unique sentence pairs across 29 language pairs.[1] Additionally, 17 corpora for other language combinations have been released as bonus corpora.

In the following sections, we review related work and focus on the human evaluation methods. We also report about extrinsic automatic evaluation experiments through machine translation. We try to analyse how human and automatic evaluation methods relate and discuss their usefulness to to answer our questions.

2 Related work

Besides ParaCrawl, there have been a number of past and recent efforts to compile parallel corpora from web-crawled content. Among the recent ones, we find, for example, WikiMatrix (Schwenk et al., 2021), CCAligned (El-Kishky et al., 2020) or OSCAR (Ortiz Suárez et al., 2019).

Many of these parallel corpora are usually evaluated through machine translation (Khayrallah and Koehn, 2018) where automatic filtering of corpora and its impact on machine translation quality has gained interest in the last years (Koehn et al., 2018,

[1] See `https://paracrawl.eu/` for a breakdown of corpus size by language.

Proceedings of the 2nd Workshop on Human Evaluation of NLP Systems (HumEval), pages 32 - 41
May 27, 2022 ©2022 Association for Computational Linguistics

2019, 2020). Some other recent work like (Caswell et al., 2021) has, in contrast, put the focus on human evaluation and recommend techniques to evaluate and improve multilingual corpora to avoid low-quality data releases.

3 Human Evaluation

Human evaluation of the corpora in ParaCrawl was done in 3 different ways depending on the version of the corpus: a) based on error annotation of parallel sentences, b) based on post-editing (PE) of the output of MT systems trained with the crawled parallel corpora and c) based on manual searches over the parallel sentences using a concordancer.

We detail each of these methods in the following subsections.

3.1 Error annotation-based evaluation

Error annotation of parallel sentences was done following ELRC guidelines as compulsory required by the project call.[2] These guidelines define a set of labels to annotate sentences following a hierarchical error typology. They literally read as follows:

1. Wrong language identification (L): means the crawler tools failed in identifying the right language.

2. Incorrect alignment (A): refers to segments having a different content due to wrong alignment.

3. Wrong tokenization (T): means the text has not been tokenized properly by the crawler tools (no separator between words).

4. MT translation (MT): refers to content identified as having been translated through a Machine Translation system. A few hints to detect if this is the case:

    - grammar errors such as gender and number agreement;
    - words that are not to be translated (trademarks for instance Nike Air => if 'Air' is translated in the target language instead of being kept unmodified);
    - inconsistencies (use of different words for referring to the same object/person);
    - translation errors showing there is no human behind.

5. Translation error refers to (E):

    - Lexical errors (omitted/added words or wrong choice of lexical item, due to misinterpretation or mistranslation),
    - Syntactic error (grammatical errors such as problems with verb tense, coreference and inflection, misinterpretation of the grammatical relationships among the words in the text).
    - Poor usage of language (awkward, unidiomatic usage of the target language and failure to use commonly recognized titles and terms). It could be due to MT translation.

6. Free translation (F): means a non-literal translation in the sense of having the content completely reformulated in one language (for editorial purposes for instance). This is a correct translation but in a different style or form. This includes figures of speech such as metaphors, anaphors, etc.

If none of these errors applied, the sentence pair should be labelled as Valid.

When more than one issue appeared in the evaluated sentences, annotators were asked to choose the first one according to the above referred error typology (1 to 6). Selecting a label was compulsory to consider the sentence evaluated and be able to complete the task, although during evaluation, if no label was selected, the sentence pair was labeled as pending.

Besides this, extra information was asked after the first evaluation campaign out of the 3 carried out to clarify some of the errors:

- Wrong language identification: whether the source, the target or both texts are wrongly identified.

- MT Translation: whether the source, the target or both text are MT-translated.

- Free translation: whether the translation should be kept, even though it is freely translated.

Moreover, after the first evaluation campaign, we asked evaluators to flag sentences which contained personal data or inappropriate language by using the check boxes on the bottom right of the screen.

[2] Seehttps://www.lr-coordination.eu/sites/default/files/common/Validation_guidelines_CEF-AT_v6.2_20180720.pdf.

3.1.1 Annotators selection and annotation tool

External annotators were selected by a language service provider (LSP). Depending on the campaign, we had 1 or 2 annotators for each language pair and between 23 and 29 language pairs. Annotators were translators and had experience in similar tasks. They were introduced to the task by the LSP project managers and received an extensive support, supervision and material from our side.

The annotation was carried out using Keops,[3] a free/open-source web-based tool to perform manual evaluation of parallel sentences. Keops covers different tasks including annotation of parallel sentences following ELRC criteria. It also supports adequacy, fluency and ranking tasks. The tool was developed inside ParaCrawl and shaped to the purpose of manual evaluation of the corpora to be released. It allows managing corpora, users, roles, projects, tasks and results.

The ELRC-based annotation screen (see figure 1) was designed to focus on a sentence pair and the annotation task itself in a user-friendly way. Annotation guidelines with examples were provided in the annotation screen to avoid users get lost. Besides this, the tool allows evaluators to navigate freely through all sentence pairs in a task, see the progress of the task, leave the task and come back at any point, access the last annotated sentence or get your own annotations or a summary in TSV format. This summary is also plotted in the results screen along with time-tracking details and a form to provide feedback on the tool.

3.1.2 Error annotation campaigns

Three error-annotation evaluation campaigns were organized for different versions of the corpora:

- Campaign 1 included 2,000 randomly sampled sentences for each of the 23 language pairs covered in ParaCrawl version 3 and 1 annotator per language pair
- Campaign 2 included 1,000 randomly sampled sentences for each of the 29 language pairs covered in ParaCrawl version 6 and 2 annotators per language pair
- Campaign 3 included 1,000 randomly sampled sentences for each of the 29 language pairs covered in ParaCrawl version 7 and 1 annotator per language pair

ParaCrawl versions 3, 6 and 7 are very different in size and in which this data was processed specially regarding alignment and cleaning components as explained in (Bañón et al., 2020).

Annotators were given 3 hours to get familiar with the project, the guidelines and the tool and to ask for doubts. They needed to complete the evaluation of 1,000 sentence pairs in 10 hours. They had a week to complete the task, once started.

They were presented the error typology and criteria in different ways: a brief oral introduction, the full guidelines in PDF, a visual help section in the annotation screen and a link to Keops Evaluator Guide[4] with examples.

Extra materials and support were provided during the evaluation campaigns when necessary: more examples and refinement of definition on error typologies, where to include issues out of the error typology, etc.

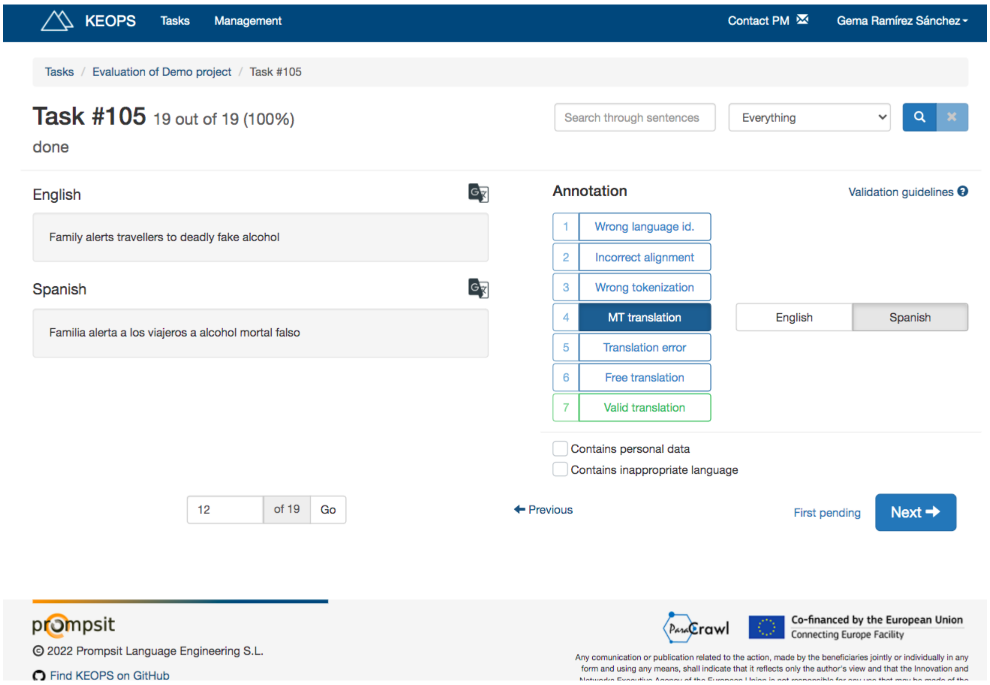
In some cases, during the course of the annotation period, we were checking actively the annotations and contacting users that were mistaken. Even though, it happened twice that we asked for a second annotator after the full task was completed because there were major issues with the 1,000 annotated sentences.

During the first evaluation campaign, we had to improvise on the fly the redefinition of some of categories to accommodate issues that were not matching any of them in the ELRC error typology that we needed to follow according to the call requirements. Namely:

- encoding issues: strange characters like Ã appeared in the texts, all due to encoding issues derived from automatic processing. We asked annotators to label those as Wrong Language.
- segmentation issues: there were sentences with partially missing text in source or target which did not match any of the categories. We asked annotators to label those as Tokenization errors.
- MT translation definition: annotators were including valid parallel sentences in this category just because they were valid but suspicious of having been produced by machine translation, we asked them not to do so but to label only bad parallel sentences that seemed to be produced by machine translation.

[3] https://github.com/paracrawl/keops

[4] https://github.com/paracrawl/keops/blob/master/evaluators.md



Task #105 19 out of 19 (100%)

done

English

Family alerts travellers to deadly fake alcohol

Spanish

Familia alerta a los viajeros a alcohol mortal falso

Annotation

1 Wrong language id.
2 Incorrect alignment
3 Wrong tokenization
4 MT translation
5 Translation error
6 Free translation
7 Valid translation

English | Spanish

□ Contains personal data
□ Contains inappropriate language

12 of 19 Go

© 2022 Prompsit Language Engineering S.L.

Any comunication or publication related to the action, made by the beneficiaries jointly or individually in any form and using any means, shall indicate that it reflects only the author's view and that the Innovation and Networks Executive Agency of the European Union is not responsible for any use that may be made of the

Figure 1: ELRC-based error annotation screen in Keops

3.1.3 Analysis of results for error annotation

Results from the first campaign were extensively reviewed by project team members. Some samples were re-annotated before determining action points on how improve the processing pipeline. We concluded that we needed better language identification, sentence segmenting or encoding fixing. But the annotation numbers themselves were considered distrustful as we observed many mislabeled sentences, mainly by lack of adherence to the hierarchy in the errors and abuse of the machine translation error category.

For example, sentences like "Hotel rooms in Paris - Habitaciones de hotel en Barcelona (Hotel rooms in Barcelona)", annotators were using MT error instead of Bad Alignment as well as for sentences like "Start your day with a good breakfast - No se puede empezar un buen día sin desayunar bien. (One cannot start a good day without a good breakfast)", very unlikely to have been produced by a MT system and probably a Free Translation.

After the first evaluation campaign, we introduced the extra information above described to be able to distinguish if the issues applied to source, target or both sides of the sentence pair or if Free translation-labelled sentences were considered as to be kept or left form the final corpus.

For the second evaluation campaign, for which we improved communication and materials about the error hierarchy adding more examples, we decided to do a second round with a second annotator. The first round results was inconclusive and even very odd for some language pairs. The second round results were very different for many languages, and, indeed, inter-annotator agreement was really low. These results are presented in table1.

For the third evaluation campaign, we tried with early spotting of annotation errors and tighter project management, but results were, again, inconclusive.

Although further annotation-based evaluation campaigns were planned in the project, we decided to replace them with other activities that could give us hints on what to focus to improve the quality of our corpora. We, though, reused the labeled sentences to perform a reassessment with the overlapping sentences from subsequent versions of the corpus.

Labelled data from all campaings is publicly available with a free/open-source licence.[5]

3.2 PE-based evaluation

When arriving at a mature phase of corpora production, and after many experiments showing that automatic metrics were improving with MT systems trained with them (see section 3 for a full explanation), we performed a PE-based evaluation

[5] https://github.com/paracrawl/human-evaluations

	L		**A**		**T**		**MT**		**E**		**F**		**V**		***IAA***
	A	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	***A-B***
Bulgarian	2	7	0	3	7	9	34	35	19	8	1	5	36	33	*0,40*
Croatian	2	1	4	4	7	5	30	23	12	11	12	2	34	53	*0,36*
Czech	3	5	36	0	8	5	17	17	3	9	1	50	31	14	*0,20*
Danish	0	0	0	0	3	0	6	58	63	5	2	0	26	37	*0,15*
Dutch	0	0	4	1	0	5	24	3	6	21	15	1	51	68	*0,22*
Estonian	0	6	5	3	10	1	48	46	17	8	0	4	19	31	*0,44*
Finnish	4	1	0	4	10	14	38	42	11	1	1	23	35	16	*0,38*
French	2	0	2	7	10	8	13	1	10	28	3	1	60	55	*0,27*
German	0	1	8	4	1	6	12	23	6	6	2	8	72	53	*0,30*
Greek	1	2	1	4	10	11	27	31	41	29	4	1	17	23	*0,42*
Hungarian	7	8	1	16	2	3	29	32	24	11	1	5	36	26	*0,41*
Icelandic	0	1	1	2	6	7	36	73	41	2	2	0	15	15	*0,23*
Irish	0	20	1	7	3	8	29	23	26	31	0	0	40	11	*0,21*
Italian	0	0	1	5	3	11	51	13	14	2	17	3	14	65	*0,15*
Latvian	5	1	1	2	8	4	26	49	26	6	2	5	32	32	*0,43*
Lithuanian	3	2	4	2	5	4	42	48	1	7	6	6	38	31	*0,47*
Maltese	0	1	4	2	19	0	51	59	2	15	1	3	23	20	*0,34*
Norwegian B.	3	5	5	10	3	4	21	0	18	28	0	16	51	36	*0,19*
Norwegian N.	1	1	24	34	0	2	1	0	9	5	8	0	57	59	*0,54*
Polish	1	0	6	3	11	1	34	50	5	8	1	5	41	33	*0,38*
Portuguese	6	3	6	6	15	3	14	5	6	1	14	2	39	78	*0,27*
Romanian	1	0	4	1	5	1	18	24	29	26	13	0	30	48	*0,16*
Slovak	3	13	3	7	3	8	27	31	14	14	14	0	36	27	*0,33*
Slovenian	5	6	4	7	8	3	46	34	12	10	6	7	18	32	*0,38*
Spanish	2	1	5	5	6	8	11	42	29	11	0	0	47	33	*0,26*
Swedish	0	1	2	7	1	5	1	19	34	21	5	9	56	39	*0,25*
Basque	0	0	7	0	0	0	15	12	53	33	2	14	23	41	-
Catalan	1	0	10	1	1	4	8	4	4	5	2	2	73	83	-
Galician	1	4	5	15	2	1	15	5	15	18	6	3	56	53	-

Table 1: Error category percentages (see error typology in section 2.1) by the two annotators (A and B) of the second evaluation campaign along with inter-annotator agreement.

experiment to have a broader view of the usefulness of our corpora to improve MT output.

To that aim, we set up an experiment to post-edit the output of the baseline MT systems and baseline + ParaCrawl MT systems created during automatic evaluation for 5 language pairs in just one translation direction (from English into 5 target languages).

3.2.1 Post-editors selection and PE tool

External post-editors were selected by an LSP to carry out the task. They were all professional translators with previous experience in PE.

The post-editing task was done using the free online MateCat CAT tool[6]. This allowed us to manage the task materials as we wanted, to invite post-editors easily and to monitor their work. MateCat makes possible the addition of user's own translation memories and also turning off any other supporting materials like machine translation or their general translation memory. In this way, we could provide the output of our systems in the form of a suggestion from a translation memory. Also for the detailed log in a spreadsheet file that we could use to perform analysis of the results.

[6] Accesible at https://www.matecat.com/

3.2.2 PE evaluation campaign

We launched just one campaign for PE-based evaluation for the final version of the corpus as the project reached its end. It was done for 1,000 words, 5 translation directions, 2 different MT systems and 3 post-editors per translation direction.

We compiled the source text to be post-edited from the online multilingual new project The Conversation[7] that publishes articles with a free/open-source licence that allows using them. We compiled the contents from a single article and segmented them while keeping the order. The article[8] was picked from a date that was out of the scope of any of the data used to train the MT systems to be evaluated.

The 15 post-editors were introduced to the tool, the details of the project, the goal of their work, etc. during a one-hour call. Instructions were shared with them also in written, and doubts were double-checked during the call:

- For every source segment, they would have two suggestions in the target language coming from two different translation memories.
- These suggestion were actually the output of machine translation but we would not tell them the particular system they were coming from.
- They needed to pick the most convenient for them to perform edits and deliver an adequate translation.
- Using external resources (dictionaries, searches, etc.) was allowed, if necessary.
- They had three days to complete the task, MateCat would track the actual time spent on it.
- In case of doubt, they should contact their project manager or ourselves.

3.2.3 Analysis of results for PE

Results (see 2) were analysed in two ways: which system was picked most frequently to perform PE and what was the edit distance (character level) from the post-edited sentence to each of the systems.

System 2 was baseline and System 1 was baseline + ParaCrawl. In all cases, the most frequently picked system was baseline + ParaCrawl.

Edit distance confirms that the final translation was closer to the output of baseline + Project-corpora than to the output of baseline. It also shows that the hardest combination to post-edit was English-Latvian, followed by English-German and English-Romanian, being English-Czech and interestingly English-Finnish the pairs with less edits. An interesting observation was that the output for baseline system for English-Czech was not so close to the baseline + Project-corpora as automatic metrics were showing in all versions of the released corpora. We deemed this information very valuable to complement the automatic evaluation based on automatic metrics only (see section 3).

3.3 Search-based evaluation

During the post-editing based campaign, we asked post-editors to use an external tool to perform searches during or after PE time.

This tool, named Corset,[9] was developed to let people perform full-index searches over the project corpora (see 2 . It also allows to select subsets of the corpora that are similar to a query document.

Internally, we had been using Corset to spot errors on the corpus looking for typical processing errors after each step in the pipeline or just doing random searches to inspect the results. This was very useful to refine the production pipeline. Also to order the results from searches on the tool based on quality heuristics.

We wanted, though, to see if professional translators found this tool useful for their work. This would give the corpora released from the project an alternative translation-related use, besides their usefulness as training data for MT.

Search-based evaluation was based on 10 manual searches, 5 language combinations a and 3 linguists per language combination.

Searchers were the same 15 professional translators working on the PE evaluation task. They were asked to perform at least 10 searches and answer a 6-question survey on their experience including usability, quality of results and value of the tool. Only 13 out of the 15 post-editors completed the work and only 11 answered the survey.

Searches were mostly related to the post-editing job content (e-bike, tyre, terrain bicycle, ubiquitous,

[7] https://theconversation.com

[8] https://theconversation.com/are-e-bikes-ruining-mountain-biking-166121

[9] https://corset.paracrawl.eu

By PE job	S1 chosen	S2 chosen	S1=S2	S1 avg ED	S2 avg ED
en-cs-nina	38	8	1	23.65	43.40
en-cs-pinta	28	15	4	40.37	53.13
en-cs-santa	32	15	1	24.18	36.77
en-de-nina	29	20	0	31.47	32.02
en-de-pinta	27	23	1	38.53	37.27
en-de-santa	30	19	0	32.43	34.39
en-fi-nina	44	7	1	30.18	52.02
en-fi-pinta	47	3	0	24.24	61.18
en-fi-santa	43	6	1	28.52	55.56
en-lv-nina	33	16	2	39.71	52.95
en-lv-pinta	32	15	1	45.65	55.76
en-lv-santa	34	13	2	46.16	58.91
en-ro-nina	39	13	1	33.84	46.43
en-ro-pinta	40	12	1	31.37	45.51
en-ro-santa	45	6	2	38.13	53.70
By language	**S1 chosen**	**S2 chosen**	**S1=S2**	**S1 avg ED**	**S2 avg ED**
en-cs	98	38	6	29.37	44.38
en-de	86	62	1	34.20	34.59
en-fi	134	16	2	27.68	56.20
en-lv	99	44	5	43.77	55.84
en-ro	124	31	4	34.44	48.55

Table 2: Post-editing (PE) results by individual jobs and by language for the most frequently chosen MT system (S1 or S2) and edit-distance (ED) from each system to the final translation

outweighs, rubbing other people's noses, mountain bikers, etc.) and a few of their own invention (medical product, disclosure statement, COVID restrictions, etc.). Most in English, and just a few in the target languages. We discovered, though, that many of the searches in English were performed on the target side of the corpus (user needs to indicate source or target) because the target side was the default option. We changed it to source after discovering so many mistaken searches.

Users reported positive feedback on the usability of the tool and the value of being able to perform searches over a parallel corpus. Some of them, though were complaining about the presence of English in the target languages, derived from the user interface mistake above mentioned. After repeating the searches setting the correct side of the corpus they were looking into, most of the negative comments turned into positive feedback about the diversity of examples and translations. Users reported also the presence of MT content and misaligned sentences in some languages.

Their feedback and our own experience showed that this simple method could be easily turned into action points although not being very systematic.

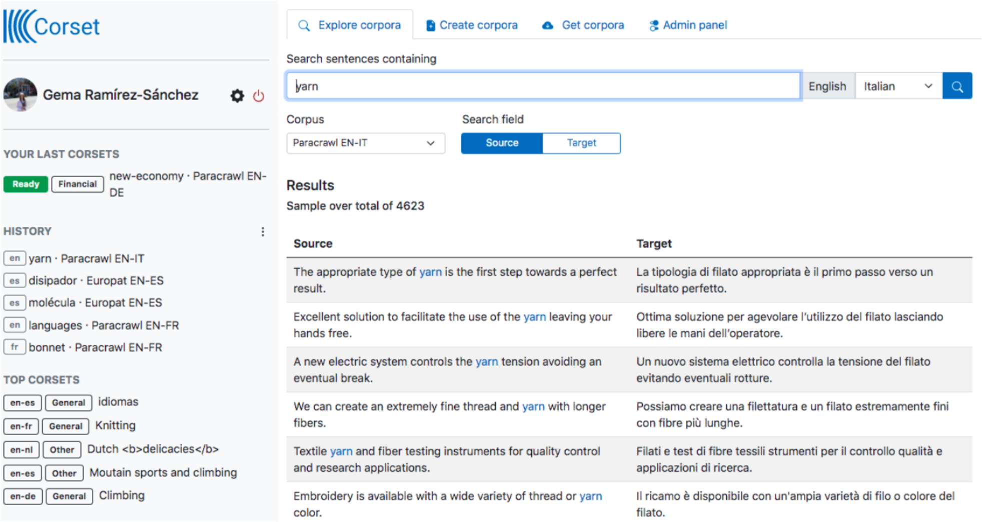
4 Automatic Evaluation

Automatic evaluation was done mainly by the addition of ParaCrawl data to WMT data from the translation shared task (Bojar et al., 2017) as an ongoing experiment carried out since the first version of the corpus released in January 2018 up to the final version until present dated from September 2021. MT evaluation based on sub samples of ParaCrawl and the addition to Europarl (Koehn, 2005) was also explored for an early version but was abandoned by lack of resources and time.

4.1 WMT-based evaluation

This experiment was designed to compare the performance of state-of-the-art neural machine translation models trained on WMT datasets (baseline) and adding ParaCrawl corpora (baseline + ParaCrawl) for five language pairs: English-Czech, English-German, English-Romanian, English-Finnish and English-Latvian

Baselines use the data from WMT17 except for English-Romanian for which the data comes from WMT16. The different ParaCrawl versions are added to WMT data to see their effect. Neural



Gema Ramírez-Sánchez

YOUR LAST CORSETS

Ready · Financial · new-economy · Paracrawl EN-DE

HISTORY

- en yarn · Paracrawl EN-IT
- es disipador · Europat EN-ES
- es molécula · Europat EN-ES
- en languages · Paracrawl EN-FR
- fr bonnet · Paracrawl EN-FR

TOP CORSETS

- en-es General idiomas
- en-fr General Knitting
- en-nl Other Dutch <b>delicacies</b>
- en-es Other Moutain sports and climbing
- en-de General Climbing

Search sentences containing: yarn · English · Italian

Corpus: Paracrawl EN-IT

Search field: Source · Target

Results

Sample over total of 4623

Source	Target
The appropriate type of yarn is the first step towards a perfect result.	La tipologia di filato appropriata è il primo passo verso un risultato perfetto.
Excellent solution to facilitate the use of the yarn leaving your hands free.	Ottima soluzione per agevolare l'utilizzo del filato lasciando libere le mani dell'operatore.
A new electric system controls the yarn tension avoiding an eventual break.	Un nuovo sistema elettrico controlla la tensione del filato evitando eventuali rotture.
We can create an extremely fine thread and yarn with longer fibers.	Possiamo creare una filettatura e un filato estremamente fini con fibre più lunghe.
Textile yarn and fiber testing instruments for quality control and research applications.	Filati e test di fibre tessili strumenti per il controllo qualità e applicazioni di ricerca.
Embroidery is available with a wide variety of thread or yarn color.	Il ricamo è disponibile con un'ampia varietà di filo o colore del filato.

Figure 2: Full-index parallel corpora search screen in Corset.

models are trained using MarianNMT (Junczys-Dowmunt et al., 2018) transformer-base with a 32,000 word SentencePiece (Kudo and Richardson, 2018) vocabulary. BLEU (Papineni et al., 2001) scores for the last four versions of the corpus systems are shown in table 3 and corpora sizes are shown in figure 4.

Further metrics such as chrF (Popović, 2015) and COMET (Rei et al., 2020) were computed. All lead to the same conclusions and even showed that version 9 of the corpus was better than 7 for English-German, contradicting BLEU. We also used a second test set, a shelf-crawled strictly multilingual TED Talks test set, for which results were all positive when adding ParaCrawl corpora to baseline with an exception for English-Czech. For this pair, the baseline was never beaten according to BLEU and chrF, in disagreement with COMET.

Comparing automatic and PE results, we noted that the little improvement in BLEU in the English-Czech baseline + ParaCrawl v9 system was having a much higher positive impact when deciding which system output to pick for PE. In all other cases, improvement in automatic metrics were higher and PE results were consistent.

Although the results show improvement for all language combinations and PE results are accordingly, there is still uncertainty about the reason of the improvement being the addition of new data more than the quality of the corpora themselves. We are also unsure about the suitability of this experiment, covering only 5 pairs, to represent the overall quality of the released corpora, which included 29 languages in its last version. Finally, we are also not convinced about the suitability of the test sets used to show the value of the corpora.

5 Conclusions and future work

We have presented in this paper a summary of the tasks carried out as part o the quality assessment activities of the ParaCrawl project to evaluate the production of web-crawled parallel corpora for machine translation. We have extensively described and discussed how we implemented different human evaluation tasks based on error annotation, post-editing and searches over the corpora and their results. We have also briefly reported about the extrinsic evaluation through machine translation conducted in parallel with human evaluation. Besides describing the methods and experiments, we have discussed their usefulness to meet the goals of the ParaCrawl project and their limitations.

The advantages and disadvantages of these methods are now being discussed in MaCoCu,[10] a similar effort for which quality assessment activities are being planned not only for bilingual corpora but also for monolingual ones. For human evaluation, annotation is probably going to be focused on single issues tasks rather that multiple and hierarchic ones. Searches and post-editing are under discussion as well as the suitability for other tasks like direct assessment, ranking and fluency, this last maybe suitable also for monolingual corpora. For extrinsic automatic evaluation, more balanced corpora sizes or not only concatenation of data but also fine tuning is being considered. Monolingual

[10] https://macocu.eu/

training corpus	cs-en	en-cs	de-en	en-de	fi-en	en-fi	lv-en	en-lv	ro-en	en-ro
WMT	28.1	21.7	33.4	27.2	24.8	21.3	18.1	15.2	33.4	28.3
WMT + PC-6	28.4	22.0	36.3	29.8	31.7	23.7	22.8	19.6	39.3	31.4
WMT + PC-7	28.0	21,9	**36.4**	30.0	32.2	24.8	23.2	19.5	39.4	31.7
WMT + PC-8	**29.0**	22.3	35.3	29.6	32.3	25.7	23.0	20.0	40.2	32.5
WMT + PC-9	**29.0**	**22.9**	36.0	**30.5**	**33.1**	**27.9**	**24.0**	**20.7**	**40.5**	**33.5**

Table 3: BLEU scores for the NMT models trained with WMT16/17 training corpora and adding ParaCrawl versions 6 to 9. Best scores are in bold.

corpus	cs	de	fi	lv	ro
WMT	52.0	5.8	2.6	4.5	0.6
PC-6	17.9	58.8	4.3	2.2	4.2
PC-7	14.0	42.8	7.3	3.7	6.2
PC-8	50.0	261.0	15.0	8.0	13.0
PC-9	50.6	278.0	31.0	13.0	25.0

Table 4: Corpus sizes in million sentences from the WMT (baseline) and ParaCrawl versions 6 to 9.

corpora will probably also be automatically tested on downstream applications or tasks.

Acknowledgements

This work has been supported by the three ParaCrawl projects (paracrawl.eu) funded by the Connecting Europe Facility of the European Union 2014-2020 - CEF Telecom, already finished and an additional ongoing project, MaCoCu (macocu.eu), also funded by the same programme under Grant Agreement No. INEA/CEF/ICT/A2020/2278341. This communication reflects only the author's view.

References

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.

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.

Isaac Caswell, Julia Kreutzer, 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 Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, 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 Balli, 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. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *CoRR*, abs/2103.12028.

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)*, pages 5960–5969, Online. Association for Computational Linguistics.

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

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Technical Report RC22176(W0109-022), IBM Research Report.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

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, Online. Association for Computational Linguistics.