Proceedings of SIGUL2022 @LREC2022, pages 59–69
Marseille, 24-25 June 2022

Quality versus Quantity: Building Catalan-English MT Resources

**Ona de Gibert, Ksenia Kharitonova, Blanca Calvo Figueras,
Jordi Armengol-Estapé, Maite Melero**
Barcelona Supercomputing Center
Plaça Eusebi Güell 1-3, Barcelona 08034, Spain
{ona.degibert, ksenia.kharitonova, blanca.calvo, jordi.armengol, maite.melero}@bsc.es

Abstract

In this work, we make the case of quality over quantity when training a MT system for a medium-to-low-resource language pair, namely Catalan-English. We compile our training corpus out of existing resources of varying quality and a new high-quality corpus. We also provide new evaluation translation datasets in three different domains. In the process of building Catalan-English parallel resources, we evaluate the impact of drastically filtering alignments in the resulting MT engines. Our results show that even when resources are limited, as in this case, it is worth filtering for quality. We further explore the cross-lingual transfer learning capabilities of the proposed model for parallel corpus filtering by applying it to other languages. All resources generated in this work are released under open license to encourage the development of language technology in Catalan.

Keywords: Machine Translation, Catalan, Under-Resourced Languages, Parallel Corpus, Data Cleaning

1. Introduction

In recent years, the arrival of the transformers (Vaswani et al., 2017) has opened up new lines of research with a clear focus on under-resourced languages (Zoph et al., 2016). The transfer-learning capabilities of pre-trained language models, such as BERT (Devlin et al., 2019), have successfully been used to solve downstream tasks employing much less task-specific annotated data. This has encouraged the development of multilingual and language-specific pre-trained language models (Martin et al., 2020). For instance, Liu et al. (2020) demonstrated that using a multilingual BART-like model (Lewis et al., 2019) for Machine Translation (MT) showed performance gains in low-resource language settings.

In the past, building MT resources has been ruled by quantity over quality, especially in low-resource scenarios, where there is little data available. In the quest for as much data as possible, large multilingual corpora such as CCAligned (El-Kishky et al., 2020), WikiMatrix (Schwenk et al., 2019) or Paracrawl (Bañón et al., 2020) are collected in mass from the web, without actively assessing their quality. The task of parallel corpus filtering aims at filtering noisy data originating from unreliable sources or misalignments to improve the quality of a bilingual dataset.

In this work, we focus on the collection and filtering of Catalan-English corpora. Despite the status of English as *lingua franca*, there are not many publicly available parallel resources for this language pair. We present new resources, both for training and evaluation, diverse in sizes and domains. Our contributions sum up to:

- A high-quality dataset for Catalan-English MT
- A quality filter for Catalan-English Parallel Corpora
- Three new evaluation datasets

Our code is openly available on Github[1] for the sake of reproducibility. We also release the resources created.

The rest of the paper is organised as follows. Section 2 provides an overview of the previous work done in the field. Section 3 describes in detail the resources presented. Section 4 outlines the human assessment of the datasets' quality. Section 5 describes our approach to the task of parallel corpus filtering. Finally, section 6 concludes our work and opens future lines of research.

2. Related Work

Typical resources to train MT models are composed of parallel corpora, i.e. bilingual aligned sentences. When trying to gather parallel training corpora for low-resource languages, a starting point is collecting large multilingual datasets, such as the ones found in OPUS (Tiedemann, 2012). OPUS includes many such datasets in a variety of languages, sizes, and domains (e.g. software handbooks, religious texts, Wikipedia articles...). Catalan is included in many of the large web-crawled corpora, however, as Kreutzer et al. (2021) point out, most data coming from online sources is of poor quality. For this reason, there is a growing interest in evaluating the quality of the released datasets. While several works focus on the quality assessment of monolingual corpora (Caswell et al., 2020), Kreutzer et al. (2021) are the first to evaluate the quality of MT datasets. They perform a large-scale human evaluation of publicly available datasets and find severe quality issues, especially for low-resource languages.

Once the quality has been assessed, a second necessary step is to improve the quality of a given dataset. Parallel corpus filtering, also known as sentence alignment filtering, is the task of automatically filtering out bad aligned sentences or sentences that are not good enough

[1] `https://github.com/TeMU-BSC/seq-to-seq-catalan/tree/main/machine_translation`

for MT training. The relevance of this task is gaining importance in recent years, as proven by the organisation of a Shared Task on Parallel Corpus Filtering and Alignment in WMT (Koehn et al., 2020).
This task has been approached using different methods that can be summarized as follows (Koehn et al., 2020):

- Filtering based on heuristic rules such as sentence length, length ratio, alpha-numerical tokens ratio, token overlap, mismatched Named Entities.
- Filtering based on automatic scores obtained by sentence embeddings or pre-trained language models.
- Filtering as a binary classification task that takes some positive and negative examples as input.

After building your model, evaluation resources are needed to test your MT system. These are much shorter in size, humanly produced, and are used as gold standards for validation. Catalan is part of the multilingual benchmark Flores-101 (Goyal et al., 2021). Other datasets for Catalan-English MT evaluation are: the Catalan United Nations test set (Costa-jussà, 2020), which is the Catalan translation of the United Nations Parallel Corpus test set (Ziemski et al., 2016), and the Catalan translation of WMT20 Biomedical Shared Task test set (Bawden et al., 2020).

3. Language Resources

In order to build a large parallel corpus for Catalan-English MT, we have compiled a total of 19 available open-source bilingual Catalan-English datasets, and we have created a brand new dataset, GEnCaTa. In total, we obtain a moderately large Catalan-English corpus of over 11.55 million aligned sentences.

3.1. Parallel Corpora Compilation

The collected datasets originate from different sources and belong to different domains. The characteristics of the corpora can be found in Table 1.
Most datasets belong to the general domain. Nonetheless, we also gather sources originating from software translations, known to contain many boilerplate sentences; Wikipedia articles, from which we expect well-constructed sentences; and specific domains, such as Health or Legislation.
We are aware that the quality of each corpus varies greatly and is difficult to measure. Furthermore, the datasets have been constructed using different methods, either produced by human translations and manual revision, or by using automatic alignment algorithms. Regarding the collected datasets, 9 out of the 20 are produced by humans.
If we look at the statistics, we can see that CCaligned contains almost as many sentences as all the other datasets together. However, it should be noted that CCaligned has been recently shown to have poor quality translations, as well as Wikimatrix (Kreutzer et al., 2021), which also has a big representation within the collected corpora for this work. Memories Lliures is the largest manually produced dataset, although its average sentences are shorter in size since it consists of a compilation of freely available translation memories, mostly coming from software. The corpora with the smallest average sentence length are Open Subtitles, movie dialogues; Tatoeba, voluntary translations; and Ubuntu, software handbooks. Not surprisingly, the longest sentences originate from Wikipedia sources, namely, Wikimedia and Wikimatrix.
To further understand the scale of the datasets, we provide a treemap visualization in the Appendix in Figure 3.

3.2. GEnCaTa: a High Quality Parallel Corpus

GEnCaTa is a high-quality Catalan-English parallel corpus composed of 38,595 segments. It has been compiled by leveraging parallel data from crawling the `gencat.cat` domain and subdomains, belonging to the Catalan Government and containing bilingual sites, both in English and Catalan.

Crawling and preprocessing We use the cleaning pipeline described in Armengol-Estapé et al. (2021) to process the WARC files obtained from the crawlings and retrieve monolingual data. Using the pipeline allows us to maintain the metadata and retrieve the original URL per each visited page.

Document alignment We extract the content of the fetched URLs from the metadata that has non-empty crawled data in both languages. We obtain 4,429 comparable sites with an average of 27.64 sentences and 382.91 and 401.65 tokens for Catalan and English, respectively. We consider each of these sites as our documents.

Sentence alignment and deduplication To align the sentences at document-level, we use the alignment algorithm Vecalign (Thompson and Koehn, 2019) based on sentence embeddings. We use multilingual sentence embeddings provided by LASER[2] for the alignment. After the automatic alignment, we obtain 126,674 aligned segments. We then perform sentence deduplication and find that almost 60% of the sentences are duplicates, leaving 51,908 parallel segments.

Manual revision A first inspection of the resulting segments has shown that the alignment was of considerable quality, which prompted us to perform a manual revision of the full dataset. Several native Catalan annotators have revised the aligned segments and labeled each pair as valid or not valid for MT training. This involves labeling as negative misaligned sentences, truncated sentences, and non-linguistic sentences.
After the manual revision of the alignment, only 38,595 segments remain (i.e. 24.98% of the aligned segments are removed).

[2] `https://github.com/facebookresearch/LASER`

	Dataset	Sentences	Tokens	Tokens/Sent	Source	Domain
1	CCaligned	5,787,682	89,606,874	15.48	(El-Kishky et al., 2020)	General
2	COVID-19 Wikipedia	1,531	34,836	22.75	(Tiedemann, 2012)	Health
3	CoVost ca-en*	263,891	809,660	10.17	(Wang et al., 2020)	General
4	CoVost en-ca*	79,633	2,953,096	11.19	(Wang et al., 2020)	General
5	Eubookshop	3,746	82,067	21.91	(Tiedemann, 2012)	Legislation
6	Europarl	1,965,734	50,417,289	25.65	(Koehn, 2005)	Legislation
7	GEnCaTa*	38,595	858,385	22.24	New	General
8	Global Voices	21,342	438,032	20.52	(Tiedemann, 2012)	General
9	Gnome*	2,183	30,228	13.85	(Tiedemann, 2012)	Software
10	JW300	97,081	1,809,252	18.64	(Agić and Vulić, 2019)	General
11	KDE4*	144,153	1,450,631	10.06	(Tiedemann, 2012)	Software
12	Memories Lliures*	1,173,055	9,452,382	8.06	Softcatalà	Software
13	Open Subtitles	427,913	2,796,350	6.53	(Lison and Tiedemann, 2016)	General
14	Opus Books	4,580	73,416	16.03	(Tiedemann, 2012)	Narrative
15	QED*	69,823	1,058,003	15.15	(Abdelali et al., 2014)	Education
16	Tatoeba*	5,500	34,872	6.34	(Tiedemann, 2012)	General
17	Tedtalks	50,979	770,774	15.12	Softcatalà	General
18	Ubuntu	6,781	33,321	4.91	(Tiedemann, 2012)	Software
19	Wikimatrix	1,205,908	28,111,517	23.31	(Schwenk et al., 2019)	Wikipedia
20	Wikimedia*	208,073	5,761,409	27.69	(Tiedemann, 2012)	Wikipedia
	Total	11,558,183	196,582,394	15.78		

Table 1: Collected parallel corpora for Catalan-English MT. *Tokens* refers to Catalan tokens. The symbol * refers to manually produced or revised datasets.

Alignment Scores We perform a further analysis of the obtained results and notice that only 19.8% of the 5,000 highest scored segments ranked by Vecalign are also selected after the manual revision. This posits the question of how much we can rely on alignment algorithms for building parallel corpora by only looking at the given score.

We release the GEnCaTa dataset with an open license, together with relevant metadata such as the source URLs and the alignment scores given by Vecalign.

4. Human Audit

As mentioned, the quality of the compiled parallel corpora differs greatly depending on domain, origin, and creation method. For that reason, as a way to uncover the unknown quality of each dataset, we follow Kreutzer et al. (2021) and perform a human evaluation of the quality of each dataset. They perform a large-scale human audit of five major multilingual datasets, including CCaligned, WikiMatrix, and Paracrawl, based on the following error taxonomy:

- **CC**: Correct translation, natural sentence
- **CS**: Correct translation, but single word or short phrase
- **CB**: Correct translation, but boilerplate
- **X**: Incorrect translation
- **WL**: Wrong language
- **NL**: Not language

They also annotate whether the segments contain offensive or porn content.

To perform our human evaluation, we randomly sample 100 aligned segments for each of the 20 datasets. Then, two native speakers conduct a blind error analysis on the 2,000 sentences, without knowing their source, and annotate each pair following the taxonomy described above.

4.1. Human Audit Results

The annotator agreement of the task obtains a score of 0.55 Cohen's Kappa, which shows moderate agreement. To compensate for the differences in human perception of the subcategories, we also report a 0.60 Kappa score considering only the binary classification correct labels (CC, CS, CB) and incorrect ones (X, WL, NL).

Results are shown in Figure 1 and in Table 7 in the Appendix. Since only 100 sentences per dataset have been evaluated, the numbers given are only rough estimates.

We combine the correct codes (CC, CB, CS) into C for simplicity. The ratio of correct samples (C) ranges from 67% to 98%. The datasets with the bigger amount of correct sentences are CoVost, sentences coming from Common Voice; Tatoeba, originating from user-generated voluntary translations; Europarl, from the European Parliament; and our brand new created GEnCaTa corpus, which is a proof of high quality. Wang et al. (2020) developed the CoVost dataset and performed data quality sanity checks based on language model perplexity, LASER scores, and a length ratio heuristic. The results of their work are in line with our findings.

On the other hand, the datasets with more mistranslations are CCAligned and Eubookshop, both originating from automatic alignments, and Ubuntu, coming from software translations.

Among the correct sentences, the corpora that contain the most boilerplate sentences are KDE4, Memories Lliures, and Ubuntu, all belonging to computer applications and handbooks. These last two also contain the

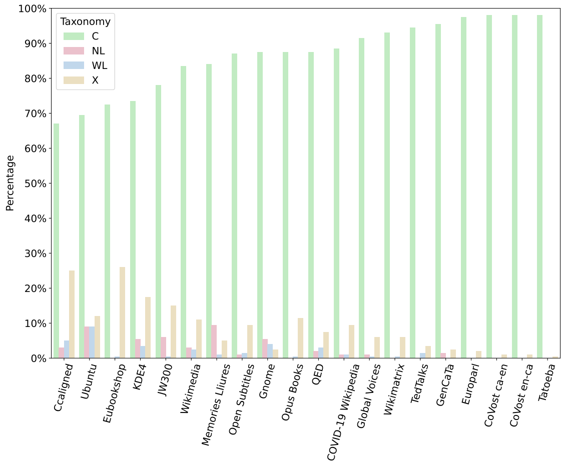

Figure 1: Results of the human audit on 20 different datasets for MT quality

biggest number of non-linguistic sentences.

The datasets containing more short sentences are Open Subtitles and Gnome, composed of dialogue and software texts, respectively.

Translations are almost always in the correct language, but it is worth to note the number of sentences in the wrong language present in Ubuntu, which refer to specific terms of computer programs.

Finally, there is no presence of offensive or porn content in most datasets, except for marginal single cases in CCAligned, CoVost, and QED.

Predictably, our analysis concludes that human revised datasets have higher quality (CoVost, GEnCaTa, Tatoeba). In the next section, we question if the effort that is needed to curate these datasets pays off.

5. Parallel Corpus filtering

Once we have compiled the parallel corpora and analysed their quality, we use the GEnCaTa dataset to build a classifier for parallel corpus filtering, by leveraging the human annotations described in Section 3.2.

5.1. Fine-tuning

Similarly to Açarçiçek et al. (2020), we fine-tune an encoder with a labeled dataset of parallel segments annotated as valid or not valid for MT. In their work, they use two small datasets of 2,000 and 10,000 samples with synthetically generated negative examples. They obtain one of the highest-performance systems in the WMT20 Shared Task on Parallel Corpus Filtering and Alignment.

We make use of the GenCaTa dataset, which consists of 51,908 samples distributed in 38,876 positive and 13,032 negative pairs. These annotations may include misaligned sentences, too short sentences, etc. We also release the labeled dataset to promote further investigations in the field. To our knowledge, this is the largest dataset available of its kind.

Label	Train	Valid	Test
Positive	23,897	7,490	7,489
Negative	8,011	2,510	2,511
Total	31,908	10,000	10,000

Table 2: Train, valid and test splits of the GEnCaTa dataset for parallel corpus filtering

Model	F1	Precision	Recall
mBERT-uncased	0.968	0.966	0.971
mBERT-cased	0.970	0.966	0.974

Table 3: Fine-tuning of mBERT results on the GEnCaTa dataset for parallel corpus filtering

We approach the task as a text classification problem and build a binary classifier that takes as input the pair of Catalan-English aligned sentences and outputs if they are valid for MT or not. Our classifier is based on mBERT (Devlin et al., 2019), a multilingual pre-trained encoder, fine-tuned with our dataset.

As shown in Table 2, we split the GEnCaTa dataset into train, valid, and test splits and then fine-tune both mBERT-cased and mBERT-uncased with the same hyperparameters. We report our results in Table 3 with almost no variability in performance but with excellent scores. We use the classifier with mBERT-cased for the subsequent experiments.

5.2. Filtering

Once we have built our classifier, we use it to filter the compiled resources described in section 3.1.

The number of total filtered sentences per dataset can be seen in Figure 2. On average, 86.87% of the original sentences are valid for MT training.

As could be expected from the human audit results, the corpora with more filtered out sentences are EUBookshop, Ubuntu, and CCaligned. Furthermore, despite the number of correct translations in Opus Books reported in the human audit, this dataset has been filtered heavily as well, since it contains quite a few misalignments and short sentences.

On the other end of the spectrum, the corpora that have been less filtered are CoVost, Tatoeba, GEnCaTa, and Europarl, the same four datasets that had the highest amount of correct sentences.

The Pearson correlation between the human audit results and the percentage of valid sentences is 0.89, a strong correlation. This proves the validity of our model, which could be used as an automatic quality estimator in the future.

5.3. Evaluation on MT Systems

We further investigate the issue of quality by assessing the impact that filtering sentence alignments may have on the quality of MT models.

For that, we build two MT systems. First, we build an MT system using the raw compiled resources (RAW).

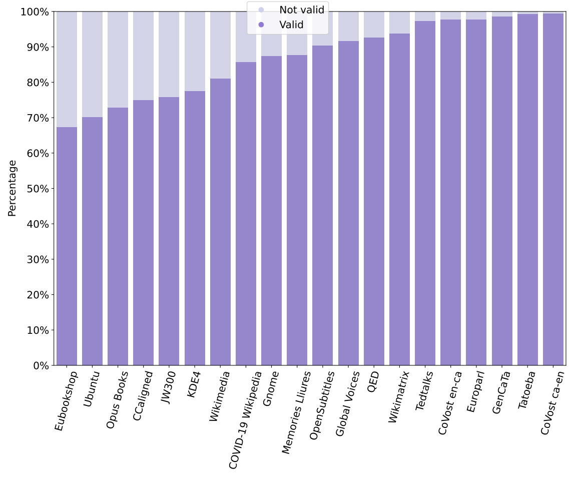

Figure 2: Percentage of filtered sentences by the parallel corpus classifier

Then, we build a new MT system to measure the impact of our parallel filtered corpus (FIL).
Both MT systems are based on mBART (Liu et al., 2020). We first pre-train a default large mBART model with concatenated monolingual data in Catalan and English and later fine-tune it with parallel data in the two languages. As monolingual data, we concatenate CaText (Armengol-Estapé et al., 2021) (in Catalan) and a clean subset of 45k random documents of Oscar (in English) (Ortiz Suárez et al., 2019).
We use default hyperparameters from Liu et al. (2020) both for monolingual pre-training and parallel data fine-tuning. However, the amount of training steps for fine-tuning is considerably lower, notably 8K (appr. 4 epochs) with an update frequency of 512. We use 4 Tesla V100-SXM2-16GB GPUs for training.
We preprocess the parallel sentences by removing duplicates, checking overlap between train and test, and removing those sentences that exceed our length limitations before feeding them to our models.

5.3.1. Evaluation Resources

We use three new in-domain test sets to validate the performance of our systems, as well as the general-domain Flores-101 as a reference. We release them under open licenses. The test sets statistics are included in Table 4.

CyberMT is a brand new test set in Catalan, Spanish, and English that belongs to the cybersecurity domain. It is composed of cybersecurity alerts extracted from the INCIBE Spanish-English corpus[3], which have been manually translated to Catalan.

TaCon is a multilingual dataset from the legal domain that includes translations of the Spanish Constitution to Basque, Catalan, Galician, Spanish, and English. To obtain it, we download the Spanish Constitution from the website of the Agencia Estatal del Boletín Oficial del Estado[4] in the corresponding languages in PDF format. We convert it to plain text, fix the broken sentences, and finally align the sentences manually.

WMT2013-ca consists of the Catalan translation of the WMT 2013 translation shared task test set (Bojar et al., 2013), belonging to the newswire domain. We commissioned the translation from Spanish to Catalan to a professional native translator.

Dataset	Languages	Domain	Sent.	Tokens
CyberMT	ca, es, en	cybersecurity	1,715	33,050
TaCon	ca, es, en eu, ga	legislation	1,110	18,275
WMT13	ca, es, en, de, ru, fr, cs	newswire	3,000	59,340

Table 4: Language resources for MT evaluation. *Tokens* refers to Catalan tokens.

5.3.2. Results

We use BLEU scores (Papineni et al., 2002) to report our results in Table 5, computed with SacreBLEU (sBLEU) (Post, 2018).

Direction	Test set	RAW	FIL
EN → CA	Cyber	40.2	**43.1**
	Flores-101	35.7	**38.0**
	TaCon	28.9	**30.2**
	WMT13	31.2	**32.9**
CA → EN	Cyber	47.4	**49.5**
	Flores-101	34.7	**37.6**
	TaCon	32.4	**35.0**
	WMT13	34.1	**36.0**

Table 5: sBLEU scores for MT evaluation

Results show that MT achieves overall good results for the Catalan-English language pair. Higher scores are obtained for the CA→EN direction, due to English being less morphologically complex.
Regarding in-domain test sets, TaCon is the test set that yields the lowest scores, probably because of the specificity of its language. Surprisingly, the Cyber test set seems to be the easiest to translate, despite being domain-specific. This may be attributed to the numerous non-verbal segments that are kept untranslated, boosting the results up to 49.5 BLEU for the CA→EN direction. Nonetheless, the most remarkable results are obtained by the comparison between the two systems. Even with the modest amount of fine-tuning steps for the two models, FIL outperforms the RAW system in all test sets. The sBLEU scores increase between 1.3 and 2.9 points. General-domain Flores-101 is the test set that shows more clearly the advantage of the quality filtering since the classifier is built on general-domain labeled data.

[3] https://www.elrc-share.eu/repository/browse/descripciones-de-vulnerabilidades-de-la-bbdd-nvd

[4] www.boe.es

	Target						
Source	CA	CS	DE	EN	ES	FR	RU
CA	-	0.952	0.979	-	0.985	0.982	0.954
CS	0.947	-	0.976	0.987	0.948	0.940	0.972
DE	0.879	0.934	-	0.987	0.937	0.949	0.958
EN	-	0.894	0.961	-	0.938	0.957	0.925
ES	0.977	0.947	0.980	0.988	-	0.982	0.971
FR	0.960	0.916	0.979	0.988	0.967	-	0.964
RU	0.936	0.972	0.979	0.981	0.975	0.969	-

Table 6: Zero-shot multilingual parallel corpus filtering

Our results show that automatically filtering sentence alignments significantly boosts MT performance and should be encouraged.

5.4. Zero-Shot Cross-lingual Transfer Learning

To further investigate the capabilities of the proposed filtering method, we explore the possibility of cross-lingual transfer learning by applying our model in zero-shot scenarios. We follow the intuition proposed by (Pires et al., 2019) that mBERT encodes multilingual representations. We use the classifier fine-tuned on CA-EN and apply it to other language pairs.

We make use of the 3,000 sentences of the WMT13 Shared Task test set for evaluation. The reason to have chosen this test set is that we have presented the Catalan version in this work, it includes six additional languages (es, en, de, ru, fr, cs) and contains document boundaries. For the synthetic test set of each language pair, we consider the 3,000 manually translated sentences as valid for MT. Then, we sample 3,000 negative examples by corrupting the alignment. To create a harder test set, we pair each sentence with the sentence of the same document that has the highest fuzzy match score to the correct translation. The final test set contains 6,000 segments.

Accuracy results are shown in Table 6. We tested all language pairs' combinations in both directions. Scores range from 0.879 to 0.988. The first insight we gain from the obtained scores is that mBERT indeed learns multilingual representations, as the results are incredibly positive. The highest scores overall are obtained by the Romance languages (CA, ES, FR); to be expected, since Spanish and French are from the same language family as Catalan. We can observe that the language typology also matters in the language direction. Since we fine-tuned the model with the direction CA→EN, the results are higher for CA and ES as a source, and for DE and EN as a target, being both Germanic languages.

Nonetheless, results are very promising for all tested combinations. While we are aware that we may introduce bias by creating a synthetic test set, we are hopeful for this new line of research that makes use of curated datasets, which may not always be available for all languages, and can later be used with new language pairs.

6. Conclusions & Future Work

In this work, we have presented the process of building high-quality MT resources for the Catalan-English language pair, which until now could be considered low-resource, and have made the case for an automatic quality filter. We have described in detail the compiled resources and the newly created ones, including a high-quality parallel corpus and three in-domain evaluation datasets. Furthermore, we have performed a human evaluation of the datasets' quality and we have devised a parallel corpus filterer, that may be used as a future quality estimator. Finally, we have applied the proposed model to zero-shot scenarios and proved the transfer-learning capabilities of mBERT.

As future lines of research, we plan to further investigate the task of quality estimation of parallel corpora and its impact on the obtained MT engines. We would also like to conduct a more qualitative analysis of the output of the MT systems to gain linguistic insights from the results.

We hope that our work encourages this line of research in the field.

7. Acknowledgements

This work was funded by the MT4All CEF project.[5]

8. Bibliographical References

Açarçiçek, H., Çolakoğlu, T., Hatipoğlu, P. E. A., Huang, C. H., and Peng, W. (2020). Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation*, pages 940–946.

Armengol-Estapé, J., Carrino, C. P., Rodriguez-Penagos, C., de Gibert Bonet, O., Armentano-Oller, C., Gonzalez-Agirre, A., Melero, M., and Villegas, M. (2021). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online, August. Association for Computational Linguistics.

Bawden, R., Di Nunzio, G., Grozea, C., Unanue, I., Yepes, A., Mah, N., Martinez, D., Névéol, A., Neves, M., Oronoz, M., et al. (2020). Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *5th Conference on Machine Translation*.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Caswell, I., Breiner, T., van Esch, D., and Bapna, A. (2020). Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus.

[5] `https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom/2019-eu-ia-0031`

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May. arXiv: 1810.04805.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., and Guzmán, F. (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.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suárez, P. O., Orife, I., Ogueji, K., Rubungo, A. N., Nguyen, T. Q., Müller, M., Müller, A., Muhammad, S. H., Muhammad, N., Mnyakeni, A., Mirzakhalov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B. F. P., Dlamini, S., de Silva, N., Çabuk Ballı, S., Biderman, S., Battisti, A., Baruwa, A., Bapna, A., Baljekar, P., Azime, I. A., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S., and Adeyemi, M. (2021). Quality at a glance: An audit of web-crawled multilingual datasets.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Seddah, D., and Sagot, B. (2020). Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, November. Association for Computational Linguistics.

9. Language Resource References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1856–1862, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., et al. (2020). Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567.

Costa-jussà, M. R. (2020). Catalan united nations v1.0 test set, June. This work is supported by the Spanish Ministerio de Economía y Competitividad and European Regional Development Fund, through the postdoctoral senior grant Ramón y Cajal.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (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)*, November.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F., and Fan, A. (2021). The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Ortiz Suárez, P. J., Sagot, B., and Romary, L. (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.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Wang, C., Wu, A., and Pino, J. (2020). Covost 2: A massively multilingual speech-to-text translation corpus.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May. European Language Resources Association (ELRA).

A. Published Resources

- The GEnCaTa Parallel Corpus
- Catalan WMT2013 MT Shared Task Test Set
- Cyber MT Test Set
- TaCon: Spanish Constitution MT Test Set
- The GEnCaTa Dataset for Parallel Corpus Filtering
- Model for English-Catalan Parallel Corpus Filtering

B. Collection of Parallel Corpora

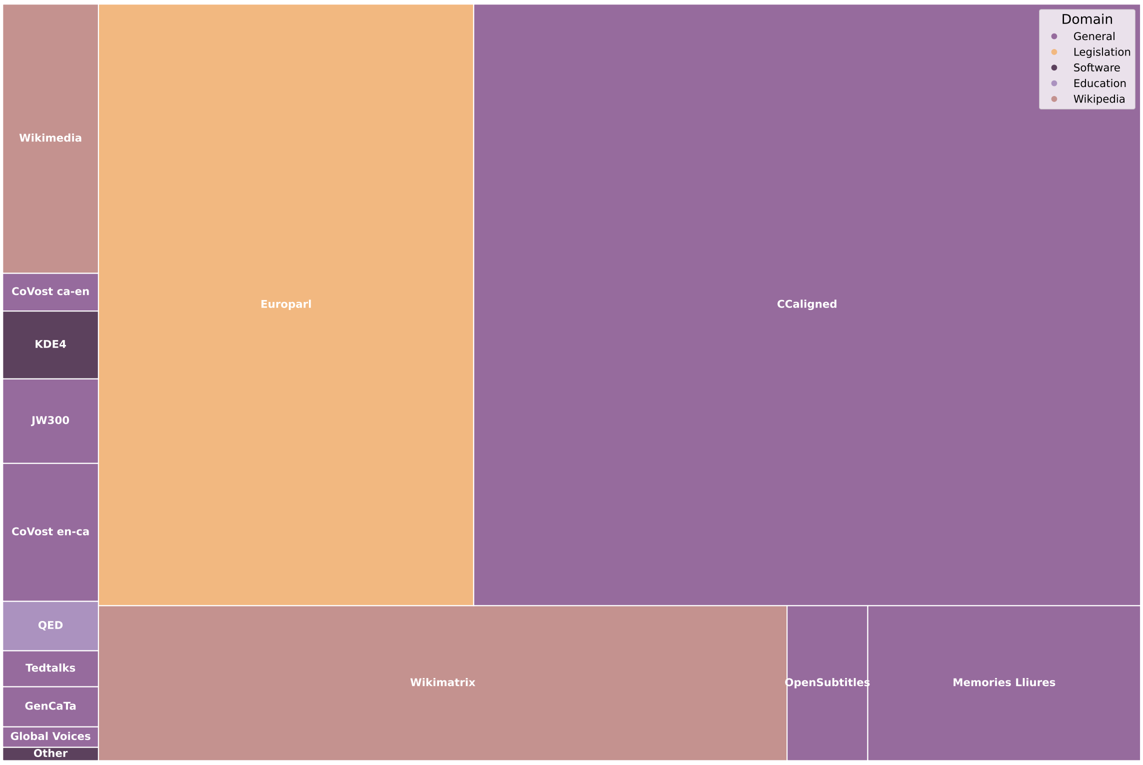

Figure 3: Treemap of the collected English-Catalan parallel corpora by number of sentences

C. Human Audit Results

Dataset	CC	CB	CS	C	X	WL	NL	offensive	porn	% audited
CCAligned	34.50%	21.00%	11.50%	67.00%	25.00%	5.00%	3.00%	0.00%	1.00%	0.0018
COVID-19 Wikipedia	82.00%	6.00%	0.50%	88.50%	9.50%	1.00%	1.00%	0.00%	0.00%	6.5317
CoVost ca-en	92.50%	2.00%	4.50%	99.00%	1.00%	0.00%	0.00%	0.00%	0.00%	0.1256
CoVost en-ca	96.00%	0.50%	2.50%	99.00%	1.00%	0.00%	0.00%	1.00%	0.00%	0.0379
GEnCaTa	79.00%	14.00%	3.00%	96.00%	2.50%	0.00%	1.50%	0.00%	0.00%	0.2592
Eubookshop	63.00%	7.50%	3.00%	73.50%	26.00%	0.50%	0.00%	0.00%	0.00%	2.6696
Europarl	96.00%	1.50%	0.50%	98.00%	2.00%	0.00%	0.00%	0.00%	0.00%	0.0051
Global Voices	77.50%	13.00%	2.00%	92.50%	6.00%	0.50%	1.00%	0.00%	0.00%	0.4686
Gnome	45.00%	25.00%	17.50%	87.50%	2.50%	4.00%	6.00%	0.00%	0.00%	4.5809
JW300	73.50%	3.50%	1.50%	78.50%	15.00%	0.50%	6.00%	0.00%	0.00%	0.1031
KDE4	19.50%	42.50%	11.50%	73.50%	17.50%	3.50%	5.50%	0.00%	0.00%	0.0694
Memories Lliures	16.00%	55.00%	13.50%	84.50%	5.00%	1.00%	9.50%	0.00%	0.00%	0.0086
Open Subtitles	66.00%	2.50%	19.50%	88.00%	9.50%	1.50%	1.00%	0.00%	0.00%	0.0234
Opus Books	74.50%	6.50%	7.00%	88.00%	11.50%	0.50%	0.00%	0.00%	0.00%	2.1835
QED	78.50%	3.00%	6.00%	87.50%	7.50%	3.00%	2.00%	1.00%	0.00%	0.1433
Tatoeba	84.00%	2.50%	13.00%	99.50%	0.50%	0.00%	0.00%	0.00%	0.00%	1.8182
Tedtalks	83.50%	3.50%	8.00%	95.00%	3.50%	1.50%	0.00%	0.00%	0.00%	0.1962
Ubuntu	13.50%	44.00%	12.00%	69.50%	12.00%	9.00%	9.50%	0.00%	0.00%	1.4748
WIkimatrix	91.50%	2.00%	0.00%	93.50%	6.00%	0.50%	0.00%	0.00%	0.00%	0.0103
Wikimedia	72.50%	7.00%	4.00%	83.50%	11.00%	2.50%	3.00%	0.00%	0.00%	0.0481

Table 7: Results of the human audit on 20 different datasets for MT quality

D. Fine-tuning Hyperparameters

D.1. Parallel Corpus Filtering

Hyper-parameter	Value
Learning Rate	0.8e-5
Learning Rate Decay	Linear
Warmup	0.06
Batch Size	64
Batch size per GPU	8
Update freq.	1
GPUs	8
Weight Decay	0.01
Max. Training Epochs	10

Table 8: Hyper-parameters used for fine-tuning the model for parallel corpus filtering. The rest of the parameters are the same as in Devlin et al. (2019)

D.2. MT training

Hyper-parameter	Value
LR scheduler	Polynomial Decay
Peak LR	1e-4
Warmup	0.2K
Total updates for LR scheduler	100K
Batch size	2048
Batch size per GPU	1
Update freq.	512
GPUs	4
Weight Decay	0.01
Max. Training Epochs	5
Dropout	0.1
Attention Dropout	0.1

Table 9: Hyper-parameters used for fine-tuning the MT models. The hyper-parameters for bilingual CA-EN denoising pre-training are the same as in Liu et al. (2020)