Universitat Pompeu Fabra Barcelona

Treball de fi de màster de Recerca

Building a Catalan-Chinese Parallel Corpus from Wikipedia for Use in Machine Translation

Nom i Cognoms Chenyue Zhou

Màster: Lingüística Teòrica i Aplicada

Edició: 2021-2022

Directors: Dra. Maite Melero Dra.

Any de defensa: 2022

Col·lecció: Treballs de fi de màster

Departament de Traducció i Ciències del Llenguatge

Acknowledgments

This work was funded by the AINA project[1].

To begin with, I would like to express my sincere gratitude to Dr. Maite Melero, the supervisor of the present thesis, for all her knowledge, patience, continuous support and time from the very beginning, for the opportunity she offered to carry out this research at Text Mining Unit, Barcelona Supercomputing Center.

This project would not have been possible without the support of many of the TEMU team members at BSC. Great thanks to Dr. Carlos Escolano, who helped conduct the data collection process which is of vital importance for the whole project. My sincere thanks go to Ona de Gilbert for her lecture on the Master's NLP course, her valuable guidance on the project plan, and of course for her "magical" data filter. I would also like to thank Ksenia Kharitonova for her great help during the project, especially for her instructions on cluster usage and the model finetuning. It has been a precious experience working with and learning from them.

Also special thanks to my teammate Zixuan Liu, for working together and encouraging each other. Also, thanks to Pol Garriga Riba, who helped us a lot not only during this project but also during the Master's courses.

Last but not least, I would like to show my deepest gratitude to my family and friends for their constant emotional support, care and love.

[1] https://www.projecteaina.cat/

Abstract

The lack of parallel corpora is one of the biggest challenges hindering progress in Machine Translation for low-resource languages. In this work, we crawl and filter parallel sentences in Catalan and Chinese from Wikipedia in order to compile a parallel corpus of good quality. This paper describes the processes we follow to build the corpus, including mining the text data, computing sentence embeddings, extracting sentence alignment and filtering for better corpus quality. We manually audit the corpus quality based on an error taxonomy. Results show that the automatic filtering we applied makes a great improvement in the quality of our web-crawled corpus. The corpus is later used as training data to finetune a multilingual Machine Translation (MT) system in both CA→ZH and ZH→CA directions. Results show that finetuning with our corpus successfully managed to improve BLEU score in both directions on the Flores-101 public benchmark test sets, which demonstrates the importance of corpus in MT and the quality of our Catalan-Chinese parallel corpus.

Keywords: Parallel Corpus, Data Mining, Corpus Quality, Machine Translation, Catalan, Chinese, Low-resource languages

Table of contents

1 Introduction

1.1 Corpus Matters for Machine Translation

With the rapid development of Artificial Intelligence, Machine Translation has become one of the most important tasks since the 1950s. There are several different periods and stages of development for MT, including rule-based methods, statistical methods, and recently proposed neural network-based learning methods. Same as most of the current approaches in Natural Language Processing (NLP), Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) are also data-driven methods because they require large-scale, high-quality parallel corpus to obtain good translation results.

NMT has seen a tremendous speed of growth in less than ten years and has already entered a mature phase. NMT is now capable of outperforming SMT when having a large quantity of data available (Koehn & Knowles, 2017). While NMT has obtained breakthrough improvements in standard benchmarks, it is known to be particularly sensitive to the size and quality of the training data (Khayrallah & Koehn, 2018; Koehn & Knowles, 2017). Therefore, the great success of NMT is especially heavily dependent on large-scale parallel corpora with both good quality and quantity.

1.2 Low-resource Language Pairs

Monolingual texts are usually available in huge amounts for many topics and languages. However, multilingual resources, especially sentences in two languages that are mutual translations, are more limited, in particular when the two languages do not involve English (Schwenk et al., 2019).

Parallel corpora can be roughly divided into corpora of low-resource language pairs and corpora of high-resource language pairs. In Machine Translation, low-resource language pairs are those where the available parallel corpora are scarce or not large enough. High-resource language pairs (such as English and French) usually do not have dataset size concerns because researchers have created ample amounts of parallel corpora over the years.

However, having large amounts of parallel data remains a major challenge for many of the 7000+ low-resource languages currently in use around the world (Ranathunga et al., 2021). Therefore, the performance of NMT on low-resource language pairs remains sub-optimal compared to the high-resource counterparts, due to the unavailability of large parallel corpora.

Due to economic and social reasons, it is meaningful to automatically translate between most of these low-resource language pairs, especially for countries that have multiple official languages. In this context, in order to apply NMT in practical settings, effective approaches to mine and filter parallel corpora are very crucial, especially for building MT resources in low-resource languages. For this purpose, we propose to build a Catalan-Chinese parallel corpus from Wikipedia for use in MT.

1.3 Language Pair of Catalan and Chinese

Tasks of Machine Translation between Chinese and Catalan are highly challenging due to the scarcity of available resources. From the linguistic point of view, Chinese and Catalan present several morphological differences. Chinese is an analytical language without inflectional morphemes for tense, voice or gender, which results in a low morpheme-per-word ratio. In contrast, Catalan is a highly inflectional language, each Catalan word has at least one independent morpheme and these morphemes are mixed together without a clear boundary. Lexically, Chinese has a massive number of homonyms, which together with the lack of morphological inflections makes the lexical-semantic disambiguation towards Catalan even harder. Syntactically, both Chinese and Catalan follow the Subject-Verb-Object pattern, and this could theoretically decrease the reordering costs (Costa-Jussà et al., 2019).

Despite the considerable differences between these two languages, the development of MT for Catalan and Chinese still entails a great asset. For example, there is a substantial economic interest between Catalonia and China. The commercial relationship between these two communities is steadily growing. In 2015, Catalonia attracted 40% of the Chinese investment received by Spain (Casaburi, 2017). Additionally, from 2005 to 2050 China will be one of the countries that will generate more immigration, with the U.S. and Spain being the two main recipients[2]. In 2020, 65,048 Chinese nationals were living in Catalonia, making them the fourth-largest group of foreign residents, after Moroccans, Romanians, and Italians. The number of

[2] https://www.elperiodico.cat/ca/economia/20160102/asia-xina-activitat-importacio-exportacio-port-barcelona-contenido rs-4784599.

Chinese nationals in Catalonia has also increased its relative weight over time: in 2000, the 4,396 people of Chinese nationality accounted for 2.42% of foreign residents; by 2020, that percentage had doubled to 5.16% (Antolín & López, 2001).

1.4 Building a Catalan-Chinese Parallel Corpus

Parallel corpus mining is one of the data augmentation techniques used in MT. Data augmentation techniques usually do not alter the NMT architecture but generate data to train these neural architectures. Wikipedia is a good source of comparable corpora, it contains tons of texts on the same topic, these texts may not be direct translations of each other but may contain fragments of translation equivalents. Parallel sentences extracted from comparable corpora are considered a good source of synthetic data for MT.

To mine Catalan-Chinese parallel sentences from Wikipedia, we apply and extend an existing data mining pipeline. After sentence splitting and tokenization, we use a recently introduced open-source LASER toolkit [3] to represent the mined sentences as sentence embeddings. With this toolkit, we also compute the sentence alignment and extract the alignment with two different thresholds. The threshold values are achieved by an improved cosine similarity measurement method called margin-based scoring according to Artetxe & Schwenk (2019a). More details about the mining process can be found in Section 3.

[3] https://github.com/facebookresearch/LASER

1.5 Data Cleaning

Compared to other large-scale data sets of high resource languages, our data is relatively limited, however, even when resources are limited, it is still worth filtering for better quality. We first apply a simple deduplication process with the Linux Uniq command to remove repetition. Then we apply a more sophisticated filtering experiment with an open-source quality filter for Catalan-English parallel corpus[4] which turns out also works well for Catalan-Chinese pair. They approach the filtering task as a text classification problem and build a binary classifier that takes as input the Catalan-English aligned sentences and outputs if they are valid for MT or not. Their classifier is based on mBERT (Devlin et al., 2019), a multilingual pre-trained encoder, fine-tuned with their dataset GEnCaTa. The characteristics of the pre-filtered and filtered Catalan-Chinese parallel corpus can be found in Section 4.

1.6 Thesis's Contributions

Our study focuses on mining bitext data for low-resource languages, namely a Catalan-Chinese parallel corpus from Wikipedia. The time-consuming mining process takes a month to finish. To avoid the low quality brought by automatic crawling we apply a data filtering step using a pre-trained filter. After the filtering, based on alignment quality, only a 3.41% of the original crawled Chinese sentences is kept in the final Corpus 1.05[5] and 4.34% of the Catalan ones. For Corpus 1.10, the percentage is even lower.

[4] https://github.com/TeMU-BSC/ seq-to-seq-catalan/tree/main/machine_ translation

[5] Corpus 1.10 has a higher proportion of sentence alignment than Corpus 1.05 in general, for more explanations please see section 3.5.

After quality filtering preprocessing steps, we use this parallel corpus to finetune the multilingual language model m2m-100 in order to improve the MT performance on the Catalan-Chinese language pair, from Catalan to Chinese and from Chinese to Catalan. Our approach improves 0.3 BLEU score on the multilingual benchmark Flores-101 (Goyal et al., 2022) for Ca→Zh and 0.5 BLEU score for Zh→Ca.

We release the resulting resources of this work under open license to encourage the development of language technology in Catalan. The datasets are now free to access in HuggingFace.[6] The resources include:

- a Catalan-Chinese Wikipedia parallel corpus

- a Catalan-Chinese dataset of 400 sentences manually annotated for translation errors following (Kreutzer et al., 2022).

1.7 Thesis Structure

In the following sections, we first go through a review of related work (Section 2). Then we describe in detail how we manage to mine the language data from Wikipedia, how to align and extract parallel sentences in Catalan and Chinese, and how to improve the corpus quality (Section 3). Next, with a random sampling of a total of 400 sentence pairs, we manually evaluate the data quality of both the pre-filtered and filtered corpus (both with two thresholds). What's more, to assess the quality of the corpus we also use it to train the m2m-100 NMT model (Section 4). The paper concludes with a discussion and thoughts on future research directions (Section 5).

[6] https://huggingface.co/datasets/projecte-aina/ca_zh_wikipedia

2 Related work

In this section, we report related work regarding Chinese-Catalan language resource generation tasks, existing work on Wikipedia language data mining and work on corpus quality filtering and evaluation.

2.1 Chinese-Catalan Language Resource Generation

There is only one previous work focusing on the language pair of Chinese-Catalan (Costa-Jussà et al., 2019). This becomes the first work on the Catalan-Chinese language pair. They apply the pivotal machine translation techniques on the Transformer (Vaswani et al., 2017) for the specific case of Chinese-Catalan, with Spanish being the pivotal language. They release the first Catalan translation gold standard with all the official United Nations languages (Arabic, Chinese, French, Russian, and Spanish). This gold standard[7], which contains 4,000 sentences, is the same as the one provided for other languages in the release of the United Nations v1.0 (Ziemski et al., 2016).

There exist other works focusing on multilingual data mining tasks that include both Catalan and Chinese. WikiMatrix was presented to automatically mine parallel sentences from Wikipedia articles in 85 languages, including various dialects or under-resourced languages (Schwenk et al., 2019). In total, they extract 135M parallel sentences for 1620 different language pairs including 90,000 Catalan and Chinese parallel sentences. Another work build and open source a 7.5B training dataset that

[7] https://zenodo.org/record/3888414#.XuEEJfJS9Bw

covers thousands of language directions with supervised data, created through large-scale mining (Fan et al., 2020). Their work is aimed to create a Many-to-Many multilingual translation model that can translate directly between any pair of 100 languages, namely the m2m-100 MT model.

2.2 Wikipedia Language Data Mining

Wikipedia is probably the largest free multilingual resource on the Internet, it is arguably the largest comparable corpus. Wikipedia articles cover many diverse topics and exist in more than 300 languages. Some pages are human-translated from an existing article not necessarily from or into English. The translated pages are independently edited and thus not necessarily 100% parallel to previous articles.

Wikipedia strongly discourages the use of unedited machine translation[8], but the existence of such articles cannot be totally excluded. In general, many articles contain sentences that are mutual translations. This makes Wikipedia a very appropriate resource to mine for parallel texts for a large number of language pairs including Catalan-Chinese.

Adafre & de Rijke (2006) are among the first to harvest parallel data from Wikipedia, working on English-Dutch language pair. They investigate the potential of Wikipedia for generating parallel corpora by applying two different methods for identifying similar text across multiple languages. Their work yielded several hundreds of Dutch/English parallel sentences. In another work, the mining approach

[8] https://en.wikipedia.org/wiki/ Wikipedia: Translation

of Munteanu & Marcu (2005) was applied to extract large corpora from Wikipedia in sixteen languages (Smith et al., 2013). Elsahar et al., (2017) searched for parallel text passages in Wikipedia by comparing their named entities and time expressions. Furthermore, Aghaebrahimian (2018) proposed an approach based on bilingual BiLSTM sentence encoders to mine German, French and Persian parallel texts with English from comparable pages on Wikipedia. Linguatools developed the Wikipedia Parallel Titles Corpora which consists of aligned Wikipedia titles in 23 languages, extended with the titles' redirects and textlinks[9]. However, given that Wikipedia titles are usually short sentences or phrases instead of entire sentences with a subject, verb and object, it seems that only modest improvements were observed when adding this resource to the training material of NMT systems.

Finally, recent work by Schwenk et al., (2019) constructs the first dataset WikiMatrix to systematically handle all languages on Wikipedia, including low-resource languages and dialects. Given that WikiMatrix contains a large volume of sentence pairs in different languages, it can be used to train and evaluate translation systems more effectively for low-resource languages.

Their experiment was carried out using LASER, a library to calculate and use multilingual sentence embeddings. It is trained on 93 languages, written in 23 different alphabets (Artetxe & Schwenk, 2019b). The training data includes all European languages, many Asian and Indian languages such as Arabic, Persian, Hebrew, etc., as well as various minority languages and dialects.

[9] https://linguatools.org/tools/corpora/wikipedia-parallel-titles-corpora/

The architecture of LASER is a sequence-to-sequence system trained on many language pairs at once with a shared joint 40k BPE vocabulary and a shared encoder consisting of a bidirectional LSTM for all languages. The sentence representation is obtained by a max-pooling operation over all encoder outputs. Then the sentence representation is fed into an LSTM decoder. In our work, we also adopted this algorithm for the Catalan-Chinese parallel corpus mining task.

2.3 Corpus Quality Evaluation

It is commonly known that datasets automatically crawled and filtered tend to have an overall lower quality than hand-curated corpus (Koehn et al., 2020). Kreutzer et al., (2022) are among the first to evaluate the quality of MT datasets. They perform a large-scale human evaluation of the quality of 205 language-specific corpora released with five major public datasets (CCAligned, ParaCrawl, WikiMatrix, OSCAR, mC4). They find that lower-resource corpora have severe quality issues. They propose solutions for effective, low-effort data auditing including an error taxonomy. Our work adopts their error taxonomy and conducts a manual quality audit for both the pre-filtered and filtered corpus.

Another most applied quality measurement is to measure the improvements the corpus brings to downstream applications (Artetxe et al., 2022), in our case, the improvement of BLEU score for Machine Translation tasks. Therefore, to look into the quality of our Catalan-Chinese parallel corpus, we perform two ways to analyze it. First, we perform a manual data audit, following the solution proposed by Kreutzer et

al., (2022). We then use our Catalan-Chinese parallel corpus to finetune an NMT model and assess the improvement of BLEU score brought by the finetuning.

3 Building the Corpus

Building NLP technologies with automatically crawled datasets is promising. This is especially true for low-resource languages because data scarcity is one of the major bottlenecks for deep learning approaches. In this section, we explain in detail how the steps are taken to mine the Catalan-Chinese parallel corpus.

To mine Catalan-Chinese parallel sentences from Wikipedia, we extend a data mining pipeline created by the UPC Machine Translation group[10]. The pipeline was intended to collect data regarding occupations in all available languages in order to conduct a gender-bias-related work. We tailored this pipeline to crawl text data of all topics only in Catalan and Chinese from Wikipedia.

3.1 Wikipedia Tree Traversal

In computer science, a tree is a widely used abstract data type that represents a hierarchical structure with a set of connected nodes. Each node in the tree can be connected to many children but must be connected to exactly one parent, except for the top root node, which has no parent. Figure 1 shows how Wikipedia pages under one category are organized as a tree with a fictitious example. “Literature” is the root node, making all nodes below its subtree, such as nodes “Definitions”, “History”, “Aesthetics”, “The influence of religious texts”, “Types of literature”, “Law” etc. “History” is the root node for “Oral literature”, “Writing”, “Early written literature”, “Publishing”, within which, node “Oral literature” is the root node for “Oratory”.

[10] This pipeline was devised by Christine Raouf and Oriol Domingo, members of MT UPC (https://mt.cs.upc.edu/people/). Their dataset is now publicly available at https://github.com/mt-upc/OccGen_dataset.

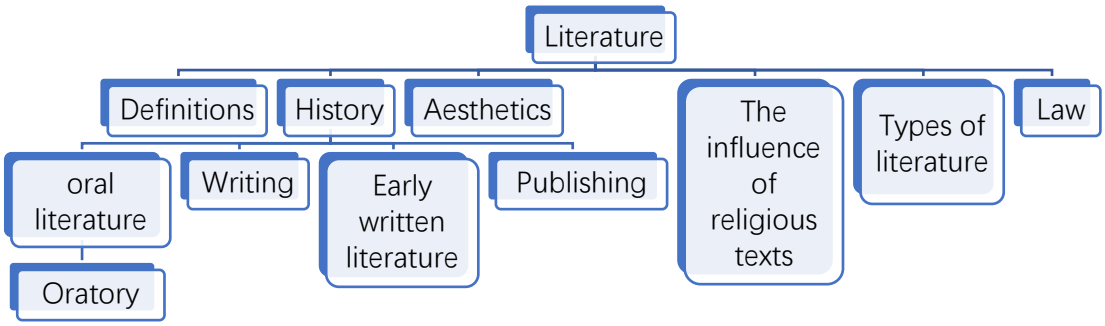

Figure 1. A fictitious example of the tree structure of Wikipedia pages.

These constraints mean there are no cycles or "loops" (no node can be its own ancestor), and also that each child can be treated as the root node of its own subtree, making recursion a useful technique for tree traversal. Tree traversal (also known as tree search and walking the tree) is a form of graph traversal and refers to the process of visiting (e.g., retrieving, updating, or deleting) each node in a tree data structure, exactly once.

For example, in our case, if we want to extract all articles related to the topic of “Literature”, we need to iterate over all nodes in the Literature tree from top to bottom. To be more specific, from the root “Literature” to the first subtree node “Definitions”, then to the second subtree node “History”, then to the third subtree node “Aesthetics” etc., meanwhile the search tree is deepened as much as possible before going to the next sibling. We set the search depth to 10 times, so as long as the visiting parent node is not empty, the search will keep going down for a maximum of 10 levels in depth. The specific time depends on the subtree levels. If the X node is empty, the search will stop once finishing visiting the X-1 node.

An internal link is a type of hyperlink on a web page to another page or resource,

such as an image or document, on the same website or domain. Internal linking allows for good website navigation and structure and allows search engines to crawl or spider websites, so Wikipedia users can navigate between pages of the same topic in different languages. In our case, we need to collect articles on the same topic both in Catalan and Chinese. Catalan Wikipedia has 702,409 articles and Chinese Wikipedia has 1,284,644 articles, so we start crawling with Catalan pages because it was easier that the Chinese version existed, than trying the other way round. By this mean, taking advantage of internal links, we collect the Chinese counterparts as well.

3.2 Processing of the Pages

We deduplicate the same articles under one category in order to increase the crawling speed. In total, we crawl and download 94,836 article pairs in Catalan and Chinese. Once we got the Catalan and Chinese articles, we use Stanza[11] (Qi et al., 2020), an open-source Python natural language processing toolkit supporting 66 human languages, to split sentences and use Google's Compact Language Detector v3 (CLD3)[12] to remove sentences that are not in Catalan or Chinese. After the sentence splitting, we have a collection of Catalan and Chinese sentences that could be potential translations. The number of the total mined and split Catalan sentences is 2,531,778 and for Chinese is 3,225,344.

[11] https://github.com/stanfordnlp/stanza
[12] https://github.com/google/cld3

3.3 Computing Sentence Embeddings

Mining parallel data consists of searching for sentences that could be potential translations in large monolingual corpora. With our collection of Catalan and Chinese monolingual sentences, we now aim to search for parallel sentences. The underlying idea of the search is to first learn a multilingual sentence embedding. In an embedding space, semantically similar sentences are close to each other no matter which language they are written in. In this scenario, the distance between two sentences in the semantic space can serve as an indicator of whether the two sentences are mutual translations or not. To compute sentence embeddings, we choose to use the freely available LASER toolkit[13]. Before BPE segmentation, we tokenize Chinese input using Jieba[14] for word segmentation.

3.4 Computing Sentence Alignment

Given two comparable corpora in two languages, the alignment task consists in identifying sentence pairs that are translations of each other. Previous works on this task usually score sentence pairs by computing the cosine similarity of their respective embeddings, then parallel sentences can be extracted through nearest-neighbor and filtered by setting a fixed threshold over this score. However, it was recently found that this approach deals poorly with scale inconsistency issues (Guo et al., 2018). So we adopted the score solution proposed by Artetxe & Schwenk (2019a) in order to address this problem. Following their algorithm, we still use LASER to compute the

[13] https://github.com/facebookresearch/ LASER
[14] https://github.com/fxsjy/ jieba

margin between the cosine of a given candidate and the average cosine of its k (Unless otherwise indicated, k = 4.) nearest neighbors in both directions as follows:

$$\text{score}(x, y) = \text{margin}(\cos(x, y), \sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2k})$$

In the formula, x and y represent the source and target sentences, and $\text{NN}_k(x)$ denotes the k nearest neighbors of x in the other language.

3.5 Extracting Alignment with Two Thresholds

One can specify the threshold on the margin score when extracting sentence alignment. In general, the higher the threshold is, the more likely the sentences are mutual translations, and the better quality the parallel sentences have. The drawback of a higher threshold is that one will get less sentences which at the same time tend to be shorter. The threshold of 1.05 is the value that Facebook recommends for LASER, so we applied this value first to extract alignment. However, when examining the sentences extracted, we found a considerable proportion of misaligned sentences, which means they are not the translation of each other. So, we then applied the threshold of 1.10 to increase the quality, meanwhile maintaining enough data to later train an NMT model.

Once the aligned sentences are extracted, we sort the corpus and apply a simple deduplication operation by Uniq command so no repeated sentences appear in the corpus. We also switched all the Traditional Chinese characters to Simplified Chinese characters to ensure the unification and to improve the later training effect on the

language model. Table 1 summarizes the statistics of the corpus extracted with the threshold 1.05 and the other one with 1.10.

Table 1. Statistics of the Corpus 1.05 and Corpus 1.10. The higher threshold leads to less data but better quality, see more details about quality in Section 4.

	Corpus 1.05		**Corpus 1.10**	
	line	**word**	**line**	**word**
CA with repetition	409,791	7,731,759	227,904	3,506,565
CA unique lines	344,715	7,589,618	169,129	3,378,652
ZH with repetition	409,791	412,246	227,904	228,958
ZH unique lines	344,715	347,167	169,129	170,183

Note: ZH words are not segmented

3.6 Quality Filtering

As Kreutzer et al., (2022) point out, most data coming from online sources is of poor quality. Even though compared to other existing large corpora, the size of our Catalan-Chinese parallel corpus is relatively limited, it is still worth filtering for quality improvement. The parallel corpus filtering task is also known as sentence alignment filtering, it is the task of automatically filtering out noisy data or misalignments or sentences that are not good for MT training. In our work, we applied the quality filter of de Gibert et al., (2022) to improve the quality of the Catalan-Chinese dataset. They approach the filtering 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. Their classifier is based on mBERT (Devlin et al., 2019), a multilingual pretrained encoder, fine-tuned with their dataset GEnCaTa.

Table 2 summarizes the statistics of the filtered corpus. The filtering shows that

67,54% of the Corpus 1.05 and 59,32% of the Corpus 1.10 are deemed invalid for MT training, which means only 32,46% of Corpus 1.05 and 40,68% of Corpus 1.10 end up staying valid for training a MT system. Given the time (one month) we devoted to the mining work, we can see that parallel data mining is a time-consuming hard work. In order to examine whether the filter does improve the corpus quality or not, we conduct a manual revision of both the pre-filtered and filtered corpus based on an error taxonomy, the audit results in Table 3 show that the filter does improve the corpus quality (Section 4).

Table 2. Statistics of the filtered corpus. It is reasonable that Corpus 1.10 has a higher proportion of aligned sentences since it has a higher threshold.

	Corpus 1.05	**Corpus 1.10**
Total	340,001	160,001
Aligned	110,360	65,084
Not aligned	229,641	94,917
Aligned %	**32,46%**	**40,68%**
Not aligned %	67,54%	59,32%

4 Corpus Quality Analysis

The size of the corpus for use in MT training is usually the primary concern, but the quality and the variety of domains may also be equally important. In our case, the source of our data satisfies the domain need, providing a variety of topics, so we turn to apply a corpus quality evaluation process.

4.1 Manual Revision

To uncover our corpus's unknown quality, we follow the evaluation method of Kreutzer et al., (2022) and perform a human audit of the quality of the corpus. We manually audit our corpora following the error taxonomy they create (provided with verbal notes on the error codes):

CC: Correct translation, natural sentence: The sentence should at least exceed five words. There's not a gold standard for the translation.

CS: Correct translation, but a single word or short phrase: Sentences shorter than five words.

CB: Correct translation, but boilerplate: Auto-generated or formulaic content or "technically correct but generally not very useful to MT models" sentences. Whether the sentence is boilerplate or not depends on subjective judgment.

C: the summary of all correct translations (CC, CS and CB).

X: Incorrect translation: Both the source and target languages are in the correct language (Catalan or Chinese), but the translations are not good.

WL: Wrong language At least one source and target are not Catalan or Chinese.

NL: Not language At least one source and target are not linguistic content.

We also annotate whether the content is offensive or pornographic. For each corpus, namely the pre-filtered Corpus 1.05, pre-filtered Corpus 1.10, filtered Corpus 1.05 and filtered Corpus 1.10, we randomly sample 100 sentence pairs and apply the human revision according to the abovementioned error taxonomy. The revision results are in Table 3. Results show that before filtering, the corpus contains a lot of X (wrong translation), most of the them are caused by misalignment. After automatic filtering, the C (summary of all correct translations) proportion increased considerably, which means the corpus quality does improve.

Table 3. Manual revision of quality for both the pre-filtered and filtered corpus. C stands for the summary of all correct translations (CC, CS and CB). The automatic filtering does improve the corpus quality.

	pre-filtered		**filtered**	
	Corpus 1.05	**Corpus 1.10**	**Corpus 1.05**	**Corpus 1.10**
Sentence pairs	340,001	160,001	110,360	65,084
C	**1%**	**4%**	**77%**	**87%**
CC	1%	3%	69%	69%
CS	0%	1%	7%	14%
CB	0%	0%	1%	4%
X	99%	96%	22%	11%
WL	0%	0%	1%	2%
NL	0%	0%	0%	0%
offensive	0%	0%	0%	0%
porn	0%	0%	1%	0%
audited %	0.03%	0.06%	0.09%	0.15%

4.2 Finetuning on NMT system

To assess the quality of an automatically crawled corpus, we can also use it as training data to measure the improvements the corpus brings to downstream applications (Artetxe et al., 2022). For Machine Translation tasks, one of the most popular evaluation metrics is BLEU (Papineni et al., 2001). So apart from the manual revision, we further investigate the issue of quality by assessing the impact our corpus may have in the performance of MT models. For that we use the m2m-100, a multilingual translation model that can translate between the 9,900 directions of 100 languages (Fan et al., 2020). We use the general-domain test sets Flores-101 (Goyal et al., 2022) to validate the finetuned m2m-100 model performance. We use BLEU scores to report the results in Table 4, computed with Sacrebleu (Post, 2018). Readers can refer to Liu (2022) for more details about the model finetuning with our corpus.

Table 4. sBLEU scores of finetuned m2m-100 model performance on Flores-101 test sets. The model is finetuned separately with the two filtered corpora.

	CA→ZH	**ZH→CA**
baseline	24.6	17.5
finetuned with filtered corpus 1.10	24.7	17.7
finetuned with filtered corpus 1.05	**24.9**	**18**

5 Conclussions and future work

In this work, we describe in detail the process of building a Catalan-Chinese parallel corpus from Wikipedia. The goal of building this corpus is to generate language resources for the Catalan-Chinese language pair, which belongs to low-resource MT tasks. By releasing our corpus, we hope to facilitate future work in Catalan NLP, and encourage open and reproducible science using public resources.

For quality examination, we perform a human evaluation and found that corpus compiled with a higher extracting threshold has a better quality in terms of the percentage of correct translations. Table 3 shows that no matter filtered or not, Corpus 1.10 outperforms Corpus 1.05 in the percentage of correct translations (prefiltered: 4% vs 1%; filtered: 87% vs 77%).

Our experiment also demonstrates that a powerful automatic data filter can significantly improve the corpus quality. Table 3 shows that after filtering, the percentage of correct translations of Corpus 1.05 increased from 1% to 77%, and for Corpus 1.10, from 4% to 87%.

On the other hand, the better finetuning results brought by Corpus 1.05 shows that data quantity also plays an important role in MT task. Given the limited time, it took us one month to finish mining the dataset of present quantity. For future work, we expect to scale up our mining process to get bigger Catalan-Chinese datasets of both good quality and quantity. Also, given that the amount of written text in Catalan is limited, another promising future direction would be developing effective multilingual data exploitation approaches, such as cross-lingual transfer methods.

6 Reference

Adafre, S. F., & de Rijke, M. (2006). *Finding Similar Sentences across Multiple Languages in Wikipedia*. 8.

Aghaebrahimian, A. (2018). *Deep Neural Networks at the Service of Multilingual Parallel Sentence Extraction*. 12.

Antolín, J. B., & López, A. S. (2001). *‘Els xinesos a Catalunya’ (2001) revisited*. 8.

Artetxe, M., Aldabe, I., Agerri, R., Perez-de-Viñaspre, O., & Soroa, A. (2022). *Does Corpus Quality Really Matter for Low-Resource Languages?* (arXiv:2203.08111). arXiv. http://arxiv.org/abs/2203.08111

Artetxe, M., & Schwenk, H. (2019a). Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3197–3203. https://doi.org/10.18653/v1/P19-1309

Artetxe, M., & Schwenk, H. (2019b). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. https://doi.org/10.1162/tacl_a_00288

Casaburi, I. (2017). Chinese investment trends in Europe. *In Europe*, 76.

Costa-Jussà, M. R., Casas, N., Escolano, C., & Fonollosa, J. A. R. (2019). Chinese-Catalan: A Neural Machine Translation Approach Based on Pivoting and Attention Mechanisms. *ACM Transactions on Asian and Low-Resource*

Language Information Processing, *18*(4), 1–8. https://doi.org/10.1145/3312575

de Gibert, O., Kharitonova, K., Figueras, B. C., Armengol-Estape, J., & Melero, M. (2022). *Quality versus Quantity: Building Catalan-English MT Resources*. 10.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Elsahar, H., Demidova, E., Gottschalk, S., Gravier, C., & Laforest, F. (2017). *Unsupervised Open Relation Extraction* (Vol. 10577, pp. 12–16). https://doi.org/10.1007/978-3-319-70407-4_3

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Grave, E., Auli, M., & Joulin, A. (2020). *Beyond English-Centric Multilingual Machine Translation* (arXiv:2010.11125). arXiv. http://arxiv.org/abs/2010.11125

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., & Fan, A. (2022). The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, *10*, 522–538. https://doi.org/10.1162/tacl_a_00474

Guo, S., Wang, Q., Wang, L., Wang, B., & Guo, L. (2018). *Knowledge Graph Embedding with Iterative Guidance from Soft Rules*. 8.

Khayrallah, H., & Koehn, P. (2018). On the Impact of Various Types of Noise on Neural Machine Translation. *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, 74–83. https://doi.org/10.18653/v1/W18-2709

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., & Guzmán, F. (2020). *Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment*. 17.

Koehn, P., & Knowles, R. (2017). *Six Challenges for Neural Machine Translation* (arXiv:1706.03872). arXiv. http://arxiv.org/abs/1706.03872

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., Suarez, P. O., … Adeyemi, M. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Munteanu, D. S., & Marcu, D. (2005). Improving Machine Translation Performance by Exploiting Non-Parallel Corpora. *Computational Linguistics*, *31*(4), 477–504. https://doi.org/10.1162/089120105775299168

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Post, M. (2018). *A Call for Clarity in Reporting BLEU Scores* (arXiv:1804.08771). arXiv. http://arxiv.org/abs/1804.08771

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages* (arXiv:2003.07082). arXiv. http://arxiv.org/abs/2003.07082

Ranathunga, S., Lee, E.-S. A., Skenduli, M. P., Shekhar, R., Alam, M., & Kaur, R. (2021). *Neural Machine Translation for Low-Resource Languages: A Survey* (arXiv:2106.15115). arXiv. http://arxiv.org/abs/2106.15115

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., & Guzmán, F. (2019). *WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia* (arXiv:1907.05791). arXiv. http://arxiv.org/abs/1907.05791

Smith, J. R., Quirk, C., & Toutanova, K. (2013). *Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment*. 9.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is All you Need*. 11.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (n.d.). *The United Nations Parallel Corpus v1.0*. 5.