CJaFr-v3: A Freely Available Filtered Japanese-French Aligned Corpus

Raoul Blin and Fabien Cromières

cnrscrlao, independent
blin@ehess.fr, fabien.cromieres@gmail.com

Abstract

We present a free Japanese-French parallel corpus. It includes 15M aligned segments and is obtained by compiling and filtering several existing resources. In this paper, we describe the existing resources, their quantity and quality, the filtering we applied to improve the quality of the corpus, and the content of the ready-to-use corpus. We also evaluate the usefulness of this corpus and the quality of our filtering by training and evaluating some standard MT systems with it.

1 Introduction

Until recently, there was no large Japanese-French corpus (several million bisegments or more), freely available. A fortiori, there was no ready-to-use for translation. The only corpus whose quality has been evaluated did not exceed 440K bisegments (Blin, 2021a), mainly constituted of TED corpus (Reimers and Gurevych, 2020). This placed Japanese-French among the low-resource language pairs. This was paradoxical because Japanese on the one hand and French on the other are languages that are very present in automatic processing. They are associated with other languages in very large corpora (for example, French-European languages, Japanese-English). There are also countless monolingual corpora of all sizes. Moreover, Japanese and French are languages with a large number of speakers (over 120M Japanese speakers; over 200M French speakers\(^1\)). The potential number of users of a bilingual corpus is therefore not negligible.

Thanks to the availability of relatively new crawled corpora of several million bisegments, the situation has changed significantly and made it possible to build a large corpus. Unfortunately, the crawled corpora available are more or less noisy and cannot be used as is.

Our goal in this work is to build and provide a large (15M bisegments), freely available, Japanese-French corpus, CJaFr-v3\(^2\). It is obtained by compiling and filtering existing resources, including crawled corpora. We also provide a standard split of the data to provide a ready-to-use corpus. In this way, we hope to significantly facilitate research on the translation and analysis of the Japanese-French pair.

This article is structured as follows. In section 2 and 3, we list the usable data. In section 4, we describe the filters and processes we applied to the crawled corpora to reduce the noise. In section 5, we evaluate the effect of the filtering by training two Transformer models on the crawled data. We observe that the filtered version provide a better model. In section 6, we describe the structure of the ready-to-use dataset.

2 Available corpora

Today, there are several large, freely available Japanese-French corpora. The OPUS project (Tiedemann and Nygaard, 2004) provides the largest number of corpora (about 20 corpora, \(\approx 29M\) sentences). Some of the corpora are also available on the site of their original producers (e.g. WikiMatrix (Schwenk et al., 2019) at https://ai.facebook.com/blog/wikimatrix/). The crawled corpora WikiMatrix, CCMatrix (Schwenk et al., 2021), MultiCCAligned (El-Kishky et al., 2020) are the largest (21M bisegments in all). They are aligned using English as a pivot language. OPUS also provides TED and Tatoeba\(^3\) corpora, all translated by humans. Tatoeba is not large and the sentences are relatively short. TED is not a native French or Japanese corpus. It is largely (if not entirely) translated from English into French and Japanese.

\(^1\)http://observatoire.francophonie.org/ qui-parle-francais-dans-le-monde/; 2021-01-10

\(^2\)http://crlao.ehess.fr/rblin/tajafr.php#cjafr

\(^3\)tatoeba.org
Among the corpora available in OPUS, we excluded very small corpora containing less than 0.1M sentences (e.g. EUbookshop). Indeed, given the expected benefit, the time required for correction was disproportionate. Six corpora (GNOME, KDE4, etc.) from computer documentation were excluded too. These are more often comparable corpora than aligned corpora. The vocabulary includes a large amount of English technical vocabulary. To avoid repetition of bisegments, we did not exploit any sub-corpora (e.g. XLEnt, extracted from WikiMatrix; CCAligned from Multi-CCAligned etc). Finally, QED is rather a comparable corpus. Its filtering was too complex and it did not allow us to obtain a sufficient number of bisegments. (Blin, 2018) noted that the bisegments in Open-subtitles (Lison and Tiedemann, 2016) are not correctly aligned. But above all, it is a corpus of very short dialogues which require special treatment because the two oral languages are very different.

Several small corpora of a few tens of thousands of bisegments have been added to those from OPUS. The Jibiki project\(^4\) provides the Cesselin, an old Japanese-French dictionary, published in 1940 and that has been OCR\(^5\) (Tomokiy, et al., 2018), and the ‘Status of the International Criminal Court’\(^7\).

The Tajaf project (Blin, 2021a) provides several small specialised corpora. Among them, two collections of titles (from French embassy website and press). Titles are interesting because of their very specific syntax in Japanese. Paradoxically, while they occupy a visible place in many types of texts, they are completely neglected in natural language processing and linguistics. Press headlines (natively in Japanese) are provided with two translations (litteral and non-litteral). Tajaf also provides a collection of Japanese idiomatic expressions translated into French and PUD\(^6\), used as test corpus.

### 3 Quality

The content and the form of the sub-corpora of CJaFr are very diverse. It contains oral presen-

tations (TED), familiar expressions (tatoeba), legal texts (CCI, tax convention), encyclopedic style (Wikipedia) etc. We assume that the crawled corpora are by nature very diverse and lacking uniformity.

We assume that the diversity is quantitatively perceptible through two measures at least: the variations of the segment lengths, and the richness of the vocabulary. We measure the length of the segments and the standard deviation of the length/average-length ratio. To measure and compare the richness of the vocabulary, for each sub-corpus we randomly extract the same number (\(N = 1000\)) of words and measure the vocabulary (\(V\)) for this sample. The values are normalised (\(V/N\)). Data are provided in (Tab.1).

The majority of the available corpora have the same shortcomings. First, most of them do not contain native French or Japanese texts. They are often synthetic corpora, translated from English. This is typically the case with TED. Synthetic texts are not in themselves a problem if the translation is of good quality. On the other hand, it is very possible that many expressions typical of each of the two languages, Japanese and French, are missing and that more anglicisms are present. Second, because of the alignment by segments, mostly sentences, it is not possible to know the cotext.\(^8\)

This is a problem especially for retrieving elided elements in Japanese (a pro-drop language). In (Tab.1) we indicate which corpora contain sentences that originally do not have cotext.

The crawled corpora represent the main part of the available corpora. It is necessary to indicate their specific defects.

1. Presence of repeated bisegments.
2. Unusual strings of symbols (e.g. ‘;;;;;;’).
3. Many automatically translated texts with poor quality (e.g. incorrect syntax, incomprehensible meaning, etc.). This phenomenon had already been raised by (Rupopp and van der Meer, 2015).

\(^1\)\(^As a reminder, "cotext" (not “context”) refers to the sentences that immediately precede.

\(^2\)\(^Example: 彼らは強力であり、非常に難いために非常に抵抗性的動物で有名な、しかし、ils sont forts et des animaux très résistants célèbre pour être très lent, mais, “\(\approx \) They are strong and some animals very resistant famous to be very slow, but.”

\(^4\)jibiki.fr

\(^5\)Japanese version: www1.doshisha.ac.jp/ karai/ntlaw/docs/icc.htm

French version: childernandarmedconflict.un.org/key documents/french/romestatutesofthe7.html

\(^6\)Test corpus used at CoNL 2017 shared task on parsing Universal Dependencies. lindat.mff.cuni.cz/ repository/ xm-lui/handle/11234/1-2184
4. No standardisation (e.g. encoding of characters; format of numerals etc.).

5. Neither the alignment nor the quality of translations is validated by humans.

4 Bisegment filtering and processing

We proceeded to a filtering, according to the following criteria. These criteria were decided by successive approximations, in order to obtain the best translation model. Some of the criteria are already used in filtering toolbox (see for example (Aulamo et al., 2020)).

1. Elimination of repeated bisegments.

2. Filtering by structure matches (parentheses): the number of opening and closing symbols must be identical in both segments.

3. Limitation of the length of the segments(350 bytes).

4. The size ratio between the two segments in a bisegment must not exceed 1 to 3.

5. Elimination of bisegments that contain character strings we consider as not belonging to canonical expressions (e.g. “;;;;”; more than two characters among “/!?$”, or any logical characters, iconic graphems (flags, emoticons etc.); more than 20 uppercase characters, and the same for digits.

We note that only \(\approx 60\%\) of the crawled corpus remains after filtering.

The Cesselin dictionary received a specific processing. We used only the translated examples, not the lexical entries. We have systematically eliminated examples containing incongruous characters generated by the OCR processing. The Japanese language is typical of the 19th and beginning of 20th century. It differs from the contemporary language. This is particularly true of verbal and adjectival conjugations. We have roughly modernised them. The conjugations have been replaced by the contemporary forms (ka-u-te (“buy”) \(\rightarrow\) ka-tte, etc.) old kana are modernised (e.g. ゐ \(\rightarrow\) い). However, some inappropriate forms remain and the resulting text cannot be considered completely modernised. Because of this, after pro-
cessing, this corpus is usable for translating from Japanese into French, but is not reliable for French into Japanese.

5 Evaluation of filtering

We have evaluated the effect of the filtering. We trained a model on a sample of the raw crawled corpus and a model on a sample of the filtered crawled corpus. The sample size is set to $\approx 20M$ Japanese words. Japanese is segmented using mecab (Kudo, 2006) and Unidic-cwj (Oka, 2017) dictionary. French is tokenised and truecased with the standard Moses tools. Japanese and French are segmented using SentencePiece (Kudo and Richardson, 2018) (BPE segmentation) with a vocabulary set to 8000 for both. The BPE segmentation model is trained and applied independently on each corpus sample (we then have four segmentation models).

The training is executed with OpenNMT-py.2.0.0 (Klein et al., 2017). We perform two trainings for each corpus and calculate the average.

The translation was evaluated after removing segmentation and separating the punctuation. The evaluation is carried out on 2 test corpus: PUD and ted.test. The scores used are BLEU$^{10}$ and c6+w2-avgF2 (Popovic, 2015).

The results (Tab.2) show that training over the filtered corpus leads to a globally better translation quality than using the raw corpus. However, the results are not uniform. The values are significantly better for ja$>$fr, but slightly negative for fr$>$ja.

6 Construction of CJaFr

CJaFr (version 3) is a package of corpora. The first set of data is constituted of the original corpora (including filtered crawled corpora). It already has been described in sections above. The second set of data is the ready-to-use corpus. It is structured in two parts (see also table 1): a core and an extension. The core was built by gathering corpora produced by translation, not webcrawling. The quantitatively most important part consists of TED. It also contains the legal corpora, locations, titles, Tatoeba etc. This core is classically decomposed into three sub-corpora: a fine-tuning corpus (3000 bisegments) and a test corpus (3000 bisegments). It contains two more test corpora: a sample of 3000 TED bisegments and PUD. The extension is composed of the filtered crawled corpora and the Cesselin.

7 Conclusion

To our knowledge, CJaFr corpus (v3) is the only freely available ready-to-use Japanese-French corpus of its size. The core is made of corpora translated by human. The extension part is obtained by filtering crawled corpora with simple techniques. The size of the core remains much smaller than the size of the translated (vs crawled) corpora available for English-Japanese in particular. It is also smaller than the theoretical size of human learning corpus (see a discussion in (Blin, 2021b)). However, adding all or part of the extension allows us to work comfortably with neural networks, and we hope it will facilitate new works for this language pair.

The crawled corpora are an undeniably welcome resource. However, at the end of this compilation work, we wonder about the interest of crawled corpora today, for a pair of languages like Japanese and French. On one hand, the quality of the crawled corpora is very variable. The resources are difficult to control (quality, content, copyright) and contain a large amount of poor quality machine translations, made by unknown machine translators. On the other hand, it exists reasonably good (commercial) translators. We suspect that it would be preferable to build synthetic corpora with such translators: take a native tar-

| lg      | corpus | raw BLEU | chrF | filtered$^b$ BLEU | chrF |
|---------|--------|----------|------|-------------------|------|
| j$>$f   | PUD    | 8.06     | 35.80| 1.96              | 4.21 |
| j$>$f   | ted    | 5.21     | 30.42| 1.48              | 2.13 |
| j$>$f   | (a)    | 6.63     | 33.11| 1.72              | 3.17 |
| f$>$j   | PUD    | 13.73    | 28.48| -0.84             | -1.31|
| f$>$j   | ted    | 10.33    | 20.77| 0.22              | -0.16|
| f$>$j   | (a)    | 12.03    | 24.63| -0.31             | -0.73|
| (a)     | PUD    | 10.90    | 32.14| 0.56              | 1.45 |
| (a)     | ted    | 7.77     | 25.60| 0.85              | 0.98 |
| Overall |        | 9.33     | 28.87| +0.70             | +1.22|

Table 2: Test scores ((a) both language pairs or corpora (b) difference: filtered score - raw score)

$^9$The hyperparameters are those suggested in opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model; 2021/06/01; batch size= 2048; word vec size = 256
$^{10}$Calculated using multi-BLEU www.statmt.org/wmt06/shared-task/multi-bleu.perl, default settings
get corpus and generate a synthetic source corpus. Unlike the crawled corpus where both source and target texts are not checked, the interest is that at least, the target is attested to be of good quality.

References

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 150–156, Online. Association for Computational Linguistics.

Raoul Blin. 2018. Automatic evaluation of alignments without using a gold-corpus - example with french-japanese aligned corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

Raoul Blin. 2021a. Kitajaf-v1-1.0.0. Guide.

Raoul Blin. 2021b. Neural machine translation, corpus and frugality. CoRR, abs/2101.10650.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAigned: A massive collection of cross-lingual web-document pairs. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020).

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In Proceedings of ACL 2017, System Demonstrations, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Taku Kudo. 2006. Mecab: yet another part-of-speech and morphological analyzer.

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, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018, pages 66–71. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA).

Teruaki Oka. 2017. Unidic for morphological analysis with reduced model size by review of crf feature templates. In Proceedings of Language Resource Workshop, volume 2, pages 144–153. NINJAL.

Maja Popovic. 2015. chrfs:character n-gram f-score for automatic mt evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Achim Rupopp and Jaap van der Meer. 2015. TAUS Moses MT Market Report.

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

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel & free. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Mutsuko Tomokiyo, Mathieu Mangeot-Nagata, and Christian Boitet. 2018. Analyse et classification d’exemples illustratifs dans le dictionnaire «cesbelin» en utilisant google traduction et un dictionnaire unl-uws. In 11es journées du réseau Lexicologie Terminologie Traduction.