Does partial pretranslation can improve low ressourced-languages pairs?

Raoul Blin

Does partial pretranslation can improve low ressourced-languages pairs?

R.Blin, cnrs-crlao, blin@ehess.fr

Abstract

We study the effects of a local and punctual pretranslation of the source corpus on the performance of a Transformer translation model. The pretranslations are performed at the morphological (morpheme translation), lexical (word translation) and morphosyntactic (numeral groups and dates) levels. We focus on small and medium-sized training corpora (50K $\sim$ 2.5M bisegments) and on a linguistically distant language pair (Japanese and French). We find that this type of pretranslation does not lead to significant progress. We describe the motivations of the approach, the specific difficulties of Japanese-French translation. We discuss the possible reasons for the observed underperformance.

1 Introduction

There are many techniques to improve the performance of a neural translation system without changing the size of the training corpus, without increasing the computational power, and independently of the tuning of the translation system: enriching vocabulary embedding (e.g. Ding and Duh (2018)), injecting linguistic information (e.g. Sennrich and Haddow (2016)), reordering (e.g. Kawara et al. (2021)), etc. These techniques are absolutely crucial for language pairs with few corpora, and when computing power is limited.

We propose here to apply another technique, which to our knowledge has not been studied so far with neural translation. It consists in pretranslating short segments of the source corpus. We proceed with handmade rules and vocabulary translation lists. We will observe its effects on Japanese-French, with several corpus sizes (50K-2.5M bisegments). This is indeed a language pair that remains poorly endowed with large, freely accessible and good quality corpora.

The aim of pretranslation is to reduce the linguistic distance between the two languages and to facilitate learning. The advantage is that it can be applied even with limited knowledge about the translation rules between the two languages. In addition, building the pretranslation rule by hand does not necessitate annotated resources (that maybe do not exist).

In section 2 we present the difficulties specific to French-Japanese. In section 3 we describe the type of pretranslations we have carried out. In section 3, we describe an experimental setup used to evaluate the effects of those pretranslations. We will see in section 5 that it is difficult to correlate these results with the properties of the corpora.

2 Challenges of Japanese-French Translation

Japanese and French are known to be linguistically distant languages. We present here briefly the most notable points of divergence, which may impact their joint treatment.

Japanese and French use different writing systems. Japanese uses about 2300 characters. French uses about 50 (including capitals). The two systems share only Arabic numerals and some English words written in Latin characters in Japanese texts. The sharing of vocabulary is therefore of little interest when training a translation model.

Word formation in the two languages does not correspond. French is a richly inflected language. Flexion concerns almost all parts of speech. On the contrary, Japanese morphology varies only for a part of verbs and adjectives (native Japanese lexical stratum). Japanese is also considered an agglutinative language. With the same meaning, many expressions in Japanese and French are not formed at the same level: expressions formed at the morphological level in Japanese are formed at the syntactic level in French and vice versa (see example below "seem not to want to drink"; CONJ:conjugation suffix):

no- *mita-* *kuna-* *katta-* *rou*

Proceedings of the 9th Workshop on Asian Translation, pages 82–88
October 17, 2022.

RAD	CONJ	CONJ	CONJ	CONJ
drink	want	NEG	PAST	seem

sembl-	*ait*	*ne*	*pas*	*vouloir*	*boire*
RAD	CONJ	AUX	AUX	V	V
seem	PAST	NEG	NEG	want	drink

Word order is different at several levels. In writing, Japanese is an SOV language where the order is semi-free, with possible pragmatic effects. French is a SVO language. Word order also diverges within phrases. In Japanese, in most cases, the complements (nominal, propositional) precede the head of the phrase. In French, they appear on either side of the head:

daidokoro[1] *no*[2] *ookii*[3] *teeburu*[4]
kitchen GEN large table
grande[3] *table*[4] *de*[2] *<la> cuisine*[1]
"large table of <the> kitchen"

A major source of difficulty in comparing or translating the two languages is the absence in Japanese of many components that are obligatory in French. Japanese uses few quantification marks (determiners etc.) and makes extensive use of bare nouns. A bare noun phrase will often have several possible translations in French (see also a discussion involving Japanese and English in (Bond, 2001). Many sentence components are elided in Japanese. Japanese is a pro-drop language. Unlike in French, the place of absent components is not occupied by a pronoun. In addition, titles and press headlines have a specific syntax in Japanese (Noguchi, 2002).

Constraints between distant structures exist in both languages but do not concern the same parts of speech. In French, distant words can share the same gender and number marks. This is the case of subject-verb agreement, for example. Japanese is known to use floating quantifiers consisting of a numeral and a classifier. The choice of the classifier depends on the quantified noun.

[*kami/pen*][1] *wo kitto* *san*-[*mai/bon*][1] *kau.*
{paper/pen} O cert. 3 - CL/CL[1] buy
(He) certainly buy three papers/pens.

3 Pretranslation applied here

We study 5 levels of pretranslation. The pretranslations are applied recursively (the pretranslation of corpus i is added to that of corpus $i-1$). The idea is that a single modification cannot substantially improve the translation. We must therefore study an accumulation of pretranslations. An example of sentence pretranslation is given in the table 8. C0 is the baseline corpus.

3.1 C1: Compositional structures

We pretranslate two structures whose translation is in general independent of the context: numerals and dates. Japanese numerals come in two forms: Sino-Japanese system (百万) or anglophone "Arabic" system (1,000,000). The treatment of numerals may seem anecdotal, but we found that they were unexpectedly poorly translated by the models trained on small corpora. This can be explained, among other things, by the variants of notation. Pretranslation is therefore both a translation and a kind of normalisation:

The translation of dates requires a triple processing: reordering, pretranslation of the numerals, global translation.

1910年$_{year}$3月$_{month}$3日$_{day}$
$\overset{reorder}{\rightarrow}$ 3日$_{day}$3月$_{month}$1910年$_{year}$
$\overset{transl.}{\rightarrow}$ 3 mars 1910

Ambiguous expressions are left as they are, such as 一日 which means «un jour» or «le premier (du mois)».

Choosing not to translate ambiguous expressions has disadvantages. Indeed, it is possible that some occurrences of an expression are not translated. We are not able to assess the number of cases involved, nor the effect on the performance of the translation model.

3.2 C2: Suffixes, punctuation, proper names

Affixes are translated if their translation is "relatively" regular: 主義 (*shugi*) → *isme*. In general, the linguistic segmentation of Japanese separates the suffix from the radical. But in order to get closer to the French form, which does not separate the suffix from its radical, we do not separate in Japanese either. It is left to the statistical segmentation (BPE) to separate or not. The form is systematically put in the singular.

共産$_{N}$ 主義$_{SUFF}$ («*kyōsan shugi*»)
→ 共産 *isme* → 共産*isme*

Punctuation is simplified and brought closer to that of French. This concerns mainly interrogative marks: か*?* → *?*; か。 → *?* .

Most of the changes in C1 concern the translation of proper names. We used several resources: an existing dictionary (jalexgram 0.37), Wikipedia

translations and translations available in unidic-cwj. Considering the possible segmentation errors and translation errors (in particular from the Wikipedia), we roughly filtered: the ratio <source word length / target word length> must not exceed 0.4 (in bytes). We do not translate one character-words because they are frequently ambiguous. We obtain a dictionary of 30,000 translated proper names.

It should be noted that the transcription (e.g. Hepburn: *Tōkyō*, Kunrei: *Toukyou*, other: *Tokyo*) is not unified within the corpora and within the dictionary. It is therefore possible that a translation in the dictionary does not correspond to a translation in the target corpus.

3.3 C3: Common nouns (CN)

In French, CNs are variable in number and are associated with a determiner, which does not exist in Japanese. We pretranslate using the singular and do not add a determiner. Most CNs have several translations, which depend on the context. For all occurrences of a noun, we will use a single translation, and always the same one. This is therefore a very rough pretranslation. 36,700 CNs have been translated (from Jalexgram).

3.4 C4:Verbal nouns (VN) in nominal position

Japanese VNs (e.g. *benkyō*) have two distributions. Followed by a support verb, they behave as verbal radicals (e.g. benkyō$_{vn}$ suru "litt. study do; to study"). Otherwise, they are used as CNs (e.g. "studies"). The corresponding forms in French occur with a determinant, but mostly at the singular form. We translate VNs in nominal position, in singular, without determiner. Here again, this is a rough pretranslation. 7,350 VNs were used.

4 Experiment

4.1 Corpus

We use the Cjafr-v3 (Blin and Cromières, 2022) corpus[1]. To our knowledge, this is the largest and freely available "ready to use" corpus currently available. The core contains 400K bisegments translated by humans. A majority of the bisegments are from TED (Reimers and Gurevych, 2020). We add a part of the extension of Cjafr (≈2M of bisegments). It is made of various crawled corpora.

From this corpus of 2.5M bisegments, and after preprocessing, two training corpora of 50K and 500K bisegments are randomly extracted. For all experiments, the fine-tuning and evaluation corpora always remain the same (but the preprocessing is different).

The evaluation is carried out on two test corpora: PUD (1000 bisegments) [2] and ted.test (3000 bisegments from TED corpus).

The corpus are morphologically analysed and segmented using mecab (Kudo, 2006) and the Unidic-cwj (Oka, 2017) dictionary. Some segmentation errors are corrected with ad-hoc rules (the same for all the experiments, including the baseline corpus). We apply thus a BPE segmentation (12K words for Japanese, 8K words for French; with SentencePiece (Kudo and Richardson, 2018)). The segmentation model is trained with the pretranslated train corpus. A description of the corpora is provided in Tables 3, 6 and 7).In particular, we evaluate the proximity (with BLEU) between the pretranslated corpora and the target corpus, after BPE segmentation.

4.2 Training and results

The training is executed with Opennmt-py.2.0.0 (Klein et al., 2017) [3]. the `batch_size` is set to 2048 and the `word_vec_size` is set to 256.

In order to reduce the variance of the results due to the random nature of the training, we perform three trainings for each corpus and calculate the average. Table 1 and 2 provide the BLEU scores [4]. For the evaluation, punctuation is separated. The raw scores are of course very different depending on the size of the training corpus. To compare them, we propose the proportional difference between the baseline score (A) and the score after pretranslation (B): B-A/A.

Several additional settings have been experimented but no one provided a significant difference with those described above. For the sake of place, we do not present them here. Those settings are: segmentation with shared vocabulary (BPE segmentation set to 16K words; evaluation with TER (Snover et al., 2006) and Chrf (Popovic); best result instead of average; evaluation after re-tokenisation.

[1] Download from `http://crlao.ehess.fr/rblin/ tajafr.php`

[2] Test corpus used at CoNL 2017 shared task on parsing Universal Dependencies. lindat.mff.cuni.cz/ repository/ xmlui/ handle/11234/1-2184

[3] The hyperparameters are those suggested in `opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model; 2021/06/01`

[4] Calculated using multi-BLEU `www.statmt.org/wmt06/shared-task/multi-bleu.perl`, default settings

	50	500	2M5
C0	4.20	14.27	15.75
C1	0.00%	2.22%	5.31%
C2	-18.08%	4.37%	-0.74%
C3	0.32%	4.13%	5.23%
C4	2.78%	3.62%	11.64%

Table 1: BLEU score; corpus PUD; proportional variation with respect to the baseline C0

	50	500	2M5
C0	3.05	9.53	15.64
C1	-1.97%	3.57%	-5.80%
C2	-10.18%	4.20%	-0.21%
C3	-0.55%	2.66%	-1.26%
C4	-2.84%	1.54%	-13.40%

Table 2: BLEU score; corpus ted.test; proportional variation with respect to the baseline

5 Discussion

As expected, pretranslation increases the proximity (measured in BLEU, on BPE segmented corpus) between the Japanese and French corpora (see Tab.5). For the smaller training corpora (50K bisegments), the progression is a little less than 1 point (knowing that in percentage, this represents 50%). The PUD corpus shows the most notable progress (+1.63 points). However, it should be noted that, whatever the corpus, the proximity is low, with or without pretranslation (<3.80 BLEU).

The results do not show significant progress. In some cases, there is even a deterioration. Nor is there a clear causal link between the (quantified) characteristics of the corpus and the results.

Compared to the size of the training corpus (50K, 500K or 2.5M bisegments), we observe a systematic but modest improvement for the 500K bisegment corpus. This behaviour is correlated with a very slight superiority of the vocabulary variety (tab.4). In other words, for this size of corpus and type of corpus, the greater variety of the corpus could improve the translation.

Concerning the test corpora, we observe better results for PUD. Again, in parallel, we note that the variety of vocabulary is slightly higher for PUD (tab.4). Moreover, if we observe the proportional difference of the number of words in Japanese and French (# words ja - # words fr/ # words ja) we see that PUD is close to train.2M5 (tab.7). We also observe a (very slightly) higher proximity between ja and fr (BLEU) for PUD (Tab.5).

We also repeated the C2 and C4 experiment with vocabulary sharing (word number for BPE segmentation is set at 16K). The results are slightly lower than with the separated vocabularies. This can be explained by the low number of common word strings, even after pre-translation. Vocabulary sharing does not improve the results after pre-translation.

It is difficult to establish a causal relationship between the (quantifiable) characteristics of the corpora and the results. Indeed, it can be observed that corpus features such as proximity (BLEU) or vocabulary variety are present in the base corpus CO. The pre-translation does not change anything, and even reduces these features.

6 Conclusion

Based on the assumption that proximity between languages could facilitate learning by a neural translation model, we locally pre-translated words and morphosyntactic structures in the source language. No significant results were observed. Some results have deteriorated. We tried to correlate these results with quantifiable features of the corpora but no clear causal relationships appeared. Several hypotheses are possible. Either the pre-translation is not thorough enough and more components need to be pre-translated to see a notable positive effect. In this case, a more massive intervention should be considered, or even coordinated with an intervention on the target. Or the linguistic characteristics of the two languages do not allow any progress through pre-translation. This could be confirmed by carrying out the same work on another language, for example SOV language (e.g. Basque) and/or with a poorer morphology SVO language (English). We have observed three corpus sizes. There is a slight improvement for the corpus of 500K words (*vs* 50K and 2.5M). To better understand the reasons for this behaviour, we propose to repeat the experiments with intermediate corpus sizes .

Acknowledgements

I sincerely thank Fabien Cromières for his valuable advice and comments.

References

Raoul Blin and Fabien Cromières. 2022. Cjafr-v3 : A freely available filtered japanese-french aligned corpus.

Francis Bond. 2001. *Determiners and Number in English contrasted with Japanese, as exemplified in Machine Translation*. Ph.D. thesis, University of Queensland.

Shuoyang Ding and Kevin Duh. 2018. How do source-side monolingual word embeddings impact neural machine translation? *arXiv preprint arXiv:1806.01515*.

Yuki Kawara, Chenhui Chu, and Yuki Arase. 2021. Preordering encoding on transformer for translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:644–655.

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.

Takako Noguchi. 2002. "midashi" no 'bunpo' - kaidoku heno tebiki to shomondai- ['Syntax' of the "headlines" - Problems and guidance to the reading]. *kouza nihongo kyouiku*, 38:94–124.

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. chrf:character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

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.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

A Description of the corpus and example of pre-translation

	train.2M5	val	PUD	ted.test
C0	2,527,216	2,964	1,000	2,929

Table 3: # bisegments

	train.50	train.500	train.2M5	val	PUD	ted.test
fr	30.44	31.33	30.88	21.24	27.49	24.93
C0	31.74	32.27	30.80	18.81	25.83	24.95
C1	31.11	32.48	31.01	19.20	25.93	25.07
C2	32.30	32.48	30.95	19.34	26.54	25.37
C3	31.42	32.13	31.14	20.25	26.90	25.06
C4	31.78	32.25	31.03	20.21	27.19	26.13

Table 4: Vocabulary variety (# of original words in a sample of 10K words/10K).

	train.50	train.500	train.2M5	val	PUD	ted.test
C0	2.84	2.60	2.17	0.07	0.49	0.12
C1	2.82	2.60	2.18	0.12	0.70	0.18
C2	3.13	2.87	2.46	0.42	1.43	0.44
C3	3.61	3.40	2.92	0.64	1.95	0.67
C4	3.79	3.57	3.07	0.65	2.12	0.74

Table 5: Proximity of the ja src corpora and the fr corpus; BLEU scores

	train.2M5	val	PUD	ted.test
C0	19.36	25.12	33.10	25.91
C1	19.32	25.08	33.00	25.93
C2	19.37	25.13	33.40	25.96
C3	20.04	25.93	34.51	26.84
C4	20.19	26.09	34.84	27.00

Table 6: Average length of the japanese segments, after BPE segmentation

	train.2M5	val	PUD	ted.test
	48,970,553	71,855	33,673	79,270
C0	-0.11%	3.61%	-1.70%	-4.27%
C1	-0.28%	3.47%	-1.99%	-4.20%
C2	-0.05%	3.66%	-0.82%	-4.08%
C3	3.43%	6.94%	2.49%	-0.84%
C4	4.19%	7.61%	3.45%	-0.24%

Table 7: # words; proportional variation with respect to the French target corpus

C0	_196 9 _年 _8 _月 _, _パウロ _6 _世 _法 王 _が _バ チ カン _の _法律 _から _死刑 _を _廃止 _し _, _すべて _の _犯 行 _に _対し _て _死刑 _判決 _は _取り除 か _れ _た _。
C1	_août _196 9 _, _パウロ _6 _世 _法 王 _が _バ チ カン _の _法律 _から _死刑 _を _廃止 _し _, _すべて _の _犯 行 _に _対し _て _死刑 _判決 _は _取り除 か _れ _た _。
C2	_août _196 9 _, _Paul os _6 _世 _法 王 _が _V at ican _の _法律 _から _死刑 _を _廃止 _し _, _すべて _の _犯 行 _に _対し _て _死刑 _判決 _は _取り除 か _れ _た _。
C3	_août _196 9 _, _Paul os _6 _世 _pape _が _V a tic an _の _loi _から _peine _de _mort _を _廃止 _し _, _すべて _の _crime _に _対し _て _peine _de _mort _判決 _は _取り除 か _れ _た _。
C4	_août _196 9 _, _Paul os _6 _世 _pa pe _が _V a tic an _の _loi _から _peine _de _mort _を _廃止 _し _, _すべて _の _crime _に _対し _て _peine _de _mort _jugement _は _取り除 か _れ _た _。
C4’	_août _196 9 _, _Paul os _6 _世 _pa pe _が _V a tic an _の _loi _から _peine _de _mort _を _廃止 _し _, _すべて _の _crime _に _対し _て _peine _de _mort _jugement _は _取り除 か _れ _た _。
fr	_en _août _196 9 _, _le _p ape _Paul _VI _a _reti ré _la _peine _de _mort _de _la _loi _du _V ati can _et _l _’ _a _reti rée _de _toutes _les _infra ctions _.

Table 8: Example of pretranslations and BPE segmentation; C4’ is obtained sharing the vocabulary; “In August 1969 , Pope Paul VI removed the death penalty from the Vatican law and revoked it from all offences . ”