Do Not Discard – Extracting Useful Fragments from Low-Quality Parallel Data to Improve Machine Translation

Steinþór Steingrímsson steinthor18@ru.is
Department of Computer Science, Reykjavik University, Iceland

Pintu Lohar pintu.lohar@adaptcentre.ie
ADAPT Centre, School of Computing, Dublin City University, Ireland

Hrafn Loftsson hrafn@ru.is
Department of Computer Science, Reykjavik University, Iceland

Andy Way andy.way@adaptcentre.ie
ADAPT Centre, School of Computing, Dublin City University, Ireland

Abstract

When parallel corpora are preprocessed for machine translation (MT) training, a part of the parallel data is commonly discarded and deemed non-parallel due to odd-length ratio, overlapping text in source and target sentences or failing some other form of a semantic equivalency test. For language pairs with limited parallel resources, this can be costly as in such cases modest amounts of acceptable data may be useful to help build MT systems that generate higher quality translations. In this paper, we refine parallel corpora for two language pairs, English–Bengali and English–Icelandic, by extracting sub-sentence fragments from sentence pairs that would otherwise have been discarded, in order to increase recall when compiling training data. We find that by including the fragments, translation quality of NMT systems trained on the data improves significantly when translating from English to Bengali and from English to Icelandic.

1 Introduction

Neural Machine Translation (NMT) usually exhibits good performance when trained on a large amount of good-quality bilingual sentence pairs. However, developing a good-quality NMT system for language pairs with limited resources is a challenging task. When compiling a parallel corpus, and during preprocessing for training, a significant amount of sentence pairs are commonly discarded before the data can be used to train the translation model. That may not be much of a problem for high-resource language pairs, where the training data contains sufficiently large number of sentence pairs even after discarding many of them, but language pairs with limited resources can be negatively impacted if the filtering is inaccurate, as less training data may limit the quality of the translation model.

The first stage of NMT training involves preprocessing the training data in which the text pairs go through several steps such as tokenising, filtering and byte-pair encoding (Sennrich et al., 2016). In the filtering step, all the text pairs with unusual source-target sentence-length

Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation, pages 1–13
September 5, 2023, Macau SAR, China.

ratio, extremely long sentences, absence of text for either one of the languages, or other anomalies are discarded. These can be a substantial percentage of available parallel pairs, and for language pairs that have limited resources, a number which could affect performance noticeably. Although the sentence pairs are discarded due to irregularities that can be detrimental for MT systems, they often contain a considerable amount of semantically similar segment pairs at the phrase, chunk or sub-sentence level. For example, if a source-language sentence contains 50 words and its target counterpart contains 10 words, they are likely to be discarded due to odd sentence-length ratio. However, they may contain similar information and some equivalent phrases or segments. This leads us to the two research questions we seek to answer in this paper:

1. **Can deficient training data for MT be identified and refined to be more useful?**
2. **Can data commonly discarded, when compiling or pre-processing training sets for NMT, be mined for parallel sentence pairs beneficial for training?**

In order to seek answers to these questions, we conduct two experiments. In the first one, described in Section 4, we work with English–Bengali sentence pairs from the *Samanantar* parallel corpus (Ramesh et al., 2022). We score the pairs and select a subset of the highest scoring pairs for training. The discarded sentences are then divided into subsentences and treated as a comparable corpus, which we mine for sentence pairs acceptable for training. In our second experiment, described in Section 5, we work with a subcorpus of the English–Icelandic parallel corpus *ParIce* (Barkarson and Steingrímsson, 2019), which is composed of a collection of parallel texts in a number of domains. The subcorpus we work with contains regulations and other documents published in relations with the EEA agreement. We collect all sentences that did not obtain alignments during the alignment process, as well as sentence pairs filtered out due to insufficient quality. We treat these discarded sentences as we treated the English–Bengali data, i.e. divide the sentences into subsentences and mine them for sentence pairs potentially useful for MT training.

Finally, we train multiple NMT models to assess the feasibility of the approach. Our evaluation shows that MT quality can be increased by extracting useful chunks at a sub-sentence level from data that would usually be discarded.

2 Related Work

A significant amount of research has been carried out in the area of exploiting comparable corpora for MT. Karimi et al. (2018) extracted parallel sentences from Wikipedia documents by translating documents in Persian into English, and also in the reverse direction, to extract semantically equivalent sentence pairs. Steingrímsson et al. (2021b) employed three different measures to identify and score parallel sentences from comparable corpora for English–Icelandic: Crosslingual information retrieval (CLIR) based approach (Lohar et al., 2016), LaBSE, and WAScore, a word alignment based scoring mechanism introduced in the paper. Ramesh et al. (2022) extracted parallel sentences from the web by using:

- monolingual corpora crawled from web,
- OCR to extract sentences from scanned documents,
- multilingual representation models for sentence alignment, and
- nearest neighbor searching method.

Munteanu and Marcu (2006) experimented with extracting parallel sub-sentences from comparable corpora using word alignments to link words in the source and target language and

calculate a signal value to estimate the probability of all word to word links, which they use to determine if two strings of words are parallel. Other work on sub-sentential fragment extraction include Hangya and Fraser (2019), who used bilingual word embeddings to greedily align words in partly parallel sentences, and then average the word alignment scores and weigh them using segment length to decide if a given segment pair is parallel. However, we are not aware of any work till date attempting to utilize discarded parallel training data.

Recent work on developing English–Bengali MT systems include Bal et al. (2019), who proposed approaches for translating assertive, interrogative and imperative English sentences into Bengali by analysing their sentence patterns and using different Bengali grammatical rules. Paul and Purkhyastha (2020) developed an English–Bengali NMT system for the aviation domain trained on a unique English–Bengali parallel corpus in this domain. Siddique et al. (2020) built a translation system using an encoder-decoder recurrent neural network with the help of knowledge-based context vectors for mapping English and Bengali words.

Until recently, work on English–Icelandic MT was limited to an Apertium (Forcada et al., 2011) based model (Brandt et al., 2011). The ParIce corpus was published in 2018, spurring work using statistical and neural methods for English–Icelandic MT. Jónsson et al. (2020) presented the first published work on Phrase-Based Statistical MT (PBSMT) and NMT for Icelandic and, in 2021, English–Icelandic was one of the language pairs in the shared news translation task at WMT (Akhbardeh et al., 2021).

3 Methodology and Experiments

In this work, we reexamine discarded parallel training data by segmenting it and extracting semantically equivalent bilingual segments. We then utilise parallel segments extracted from the discarded data as additional parallel training data if it can be deduced from our methods that the segments will be useful for MT training. We compare the quality of the translation output to baseline models. In the case of English–Bengali, the comparison is made to a model trained on the full Samanantar corpus and to the state-of-the-art IndicTrans model (Ramesh et al., 2022), and in the case of English–Icelandic, to a model trained on the aligned and filtered corpus, without the sentence pairs mined from discarded data.

3.1 Datasets

For our first experiments, we re-evaluate English–Bengali parallel sentence pairs from the *Samanantar* corpus (Ramesh et al., 2022), the largest publicly available parallel corpora collection for 11 Indic languages. The original English–Bengali parallel training data contains 8.52 million sentence pairs, sufficiently large for NMT training. However, when inspecting random samples from the dataset, we found that not all the sentence pairs are mutual translations, although many contain parallel sub-sentences that can be useful to acquire translation knowledge.

For our second experiment, we use the raw parallel documents used to compile the EEA subcorpus of ParIce (Barkarson and Steingrímsson, 2019), obtained from the corpus publisher. We took aside $903,692$ sentence pairs that had been aligned and accepted after filtering. We then collected all other sentences in the corpus, which had been discarded at some stage in the compilation process. Some did not obtain an alignment by the sentence alignment algorithm while others were not accepted by filters. In total, this resulted in over $833K$ discarded sentences in English and over $927K$ sentences in Icelandic.

3.2 Training and evaluation

For both language pairs (English–Bengali and English–Icelandic), we train separate NMT models for both translation directions. Fairseq (Ott et al., 2019) is used to train Transformer$_{BASE}$ models, as described in Vaswani et al. (2017), except that we use byte-pair encoding with a

```
--arch transformer
--share-all-embeddings
--dropout 0.2
--label-smoothing 0.2
--criterion label smoothed cross entropy
--weight-decay 0.0001
--optimizer adam
--adam-betas '(0.9, 0.98)'
--clip-norm 0
--lr-scheduler inverse sqrt
--warmup-updates 4000
--warmup-init-lr 1e-7
--keep-last-epochs 5
--patience 5
--skip-invalid-size-inputs-valid-test
--lr 0.0005 --stop-min-lr 1e-9
--max-tokens 16000
--fp16
```

Figure 1: Hyperparameters for all trained models.

shared vocabulary size of $32K$ and set dropout to 0.2, in line with Sennrich and Zhang (2019) whose results indicate that a more aggressive dropout than applied in the original Transformer paper leads to higher BLEU scores in low and medium resource settings. We train each model on a single A100 GPU with early stopping on validation loss with the patience set to 5 epochs, using the same setup as Ramesh et al. (2022) when they trained Transformer$_{BASE}$ models to compare against their large model. For validation we use the FLORES development set (Goyal et al., 2022) for English–Bengali and the in-domain EEA development set from the ParIce 21.10 dev/test splits (Barkarson et al., 2021), compiled from held-out documents from the same source as the ParIce corpus. All our hyperparameters are given in Figure 1.

We evaluate the models automatically using BLEU scores (Papineni et al., 2002), using the test sentences from the same datasets we used for validation. We calculate the scores using SacreBLEU (Post, 2018), for them to be reproducible and comparable. For Bengali–English, we follow the process carried out by (Ramesh et al., 2022). We use the default mteval-v12a tokenizer, but, since the SacreBLEU tokenizer does not support Bengali, we first tokenize using the IndicNLP[1] tokenizer before running SacreBLEU. SacreBLEU signatures for en→bn[2], bn→en[3] and for en→is and is→en[4] are provided in footnotes.

4 Refining an English–Bengali Corpus

We begin by calculating similarity scores for each of the $8.52M$ English–Bengali sentence pairs in the Samanantar corpus. We use LASER (Artetxe and Schwenk, 2019), LaBSE, and WAScore (Steingrímsson et al., 2021b) for scoring the sentence pairs. LASER uses a pre-trained BiLSTM encoder trained on data in 93 languages to generate scores for sentence pairs. LaBSE uses dual encoder models, with the encoding architecture following the BERT Base model, and additive margin softmax which creates a large margin around positive pairs. WAScore is word alignment based and uses CombAlign (Steingrímsson et al., 2021a), which again employs multiple word aligners to arrive at accurate word alignments. In order to remove sentences most likely to be deficient, we treat this as a candidate list extracted from comparable corpora, following the methodology described in Steingrímsson et al. (2021b), using a logistic regression classifier

[1] `https://github.com/AI4Bharat/indicnlp_catalog`

[2] SacreBLEU signature: BLEU+numrefs.1+case.mixed+tok.none+smooth.exp+version.2.2.0

[3] SacreBLEU signature: BLEU+numrefs.1+case.mixed+tok.13a+smooth.exp+version.2.2.0

[4] SacreBLEU signature: BLEU+numrefs.1+case.mixed+tok.13a+smooth.exp+version.2.2.0

Dataset	Size	en→bn		bn→en	
	(#sentence pairs $\times 10^6$)	BLEU	time	BLEU	time
Samanantar	8.52	18.1	29h27m	27.9	20h2m
S_1	5.6	19.0	14h33m	27.8	19h5m
S_2	5	19.1	15h43m	***28.5***	11h22m
S_3	4	18.9	16h32m	27.2	9h8m
S_4	3	**19.5**	7h32m	26.6	6h38m
S_5	2	18.7	5h57m	25.6	5h37m
S_6	1	17.3	1h29m	23.3	1h43m
S_7	0.5	14.9	1h6m	19.9	37m

Table 1: BLEU score for models trained on different sets of sub-selected English–Bengali data until convergence. Scores in bold are highest and significantly higher than other scores according to a bootstrap resampling test.

that considers all three scores to decide which sentence pairs to filter out. We then order the remaining sentence pairs based on LaBSE similarity score and create differently sized sets of parallel sentence pairs, with one set containing the 500 thousand highest scoring pairs (S_7), another containing the 1 million highest scoring pairs (S_6), and so on. Table 1 shows the size of the original data set and the different sets of selected data. Note that the S_1 data set represents all the 5.6 million sentence pairs that our rather lenient classifier deemed acceptable. The other sets contain a subset of the sentence pairs in S_1, as described above.

4.1 Baseline

We trained models for both translation directions on the full Samanantar dataset of $8.5M$ sentence pairs and set that as a baseline for our experiment. The models achieved 18.1 and 27.9 BLEU for en→bn and bn→en respectively (see Table 1), which is somewhat below the scores of 20.3 and 32.2 reported for IndicTrans (Ramesh et al., 2022), trained on the same data. This difference may be explained by the model size. We train Transformer$_{BASE}$ models with ≈60M parameters, while IndicTrans is a very large transformer model with ≈400M parameters.

4.2 Segment pairs for similarity measurement

We evaluate and compare the models trained on different amounts of data, with the smallest datasets having the highest scoring sentence pairs in terms of the similarity score used, and find that the BLEU score rises when sentence pairs are added, but only up to a point, when it starts

Language	Original sentence	Segments after splitting
Bengali	রাজনৈতিক শক্তি ও সামরিক বাহিনীর সম্পর্ক বিষয়ে তিনি বলেন, সরকারের উচিৎ আর্মির সঙ্গে ভালো ও সামঞ্জস্যপূর্ণ সম্পর্ক বজায় রাখা।	1. রাজনৈতিক শক্তি 2. সামরিক বাহিনীর সম্পর্ক বিষয়ে তিনি বলেন 3. সরকারের উচিৎ আর্মির সঙ্গে ভালো 4. সামঞ্জস্যপূর্ণ সম্পর্ক বজায় রাখা
English	Solvents can be gasses, liquids, or solids.	1. Solvents can be gasses 2. liquids 3. solids

Figure 2: English and Bengali segments after splitting.

Type of selection/discarding	#sentence/segment pairs
Whole pairs selected	$1.2M$
Whole Bengali and Partial English	$79K$
Whole English and Partial Bengali	$88K$
Both partial	$456K$
Discarded	$1.7M$

Table 2: Result of sub-sentential selection

going down again (see Table 1). These turning points are different for each language direction. Steingrímsson et al. (2023) show that different filtering approaches may suit different translation directions, even when working with the same parallel corpus. They speculate that this may be due to lower quality text in one language than in the other, affecting the quality of translations into that language if no special effort is put into filtering these lower quality texts out especially. More complex morphology in one language, effects of translationese or other systemic factors may also play a role. In our work, while evaluating our approaches on both language directions, we aim our data selection on translating from English and into Bengali and Icelandic.

When evaluating the Samanantar subsets, shown in Table 1, the turning point is lower for the en→bn dataset, with the highest BLEU for a subset of $3M$ sentence pairs. As we do not know whether a more fine grained turning point would be below or above the $3M$ sentence pair mark, to err on the side of caution we use the $2M$ highest scoring sentence pairs as a foundation for our final system, and investigate further all the other $3.6M$ pairs from the set of $5.6M$ approved by our classifier. We generate sub-sentential segments for each of these sentences and use comparable corpora mining approaches to find optimal sentence pairs. For that we first split up the sentences in both languages using commas and conjunctions as delimiters. In English we use "and" and "or", and "ও" and "এবং" in Bengali. Figure 2 shows examples of how the sentences can be split. From the segments we generate all possible combinations of up to six adjoining sentence parts for each language. We then pair each segment combination against all segment combinations in the other language for any given pair. This results in a total of ≈115 million pairs to be evaluated, representing the $3.6M$ sentence pairs from the parallel corpus.

We use LaBSE to estimate semantic similarity for all segment pairs. Feng et al. (2022) use the threshold 0.6 for selecting sentence pairs mined from CommonCrawl,[5] as they find pairs scoring higher than or equal to this threshold likely to be at least partial translations of each other. Partial translations are often an effect of misalignment and according to Koehn et al. (2018) including them in a training set can be detrimental to the output quality of a resulting MT system. Our aim is to reduce the number of partial translations in our training set and extract from them better mutual translations. Thus, we decide to set our threshold even higher, to 0.75. Furthermore, we proceed to find the one best segment pair created from each sentence pair, and only include that in our training set. Sometimes it comprises the whole sentence on both sides and sometimes only a part of either one or both the sentences. For almost half the sentence pairs all segment pair candidates are discarded as shown in Table 2. Using this approach, we produce $1.8M$ pairs, of which $1.2M$ were complete sentence pairs and over $600K$ containing partial sentences on either one or both sides. We add these to our foundation training set of $2M$ sentence pairs and then use this combined data to train a new translation model to investigate whether this processing approach affects the quality of translations, as measured by BLEU.

[5] `http://commoncrawl.org/`

Direction	BLEU	time
en→bn	**19.7**	10h52m
bn→en	26.8	10h32m

Table 3: BLEU scores for the final English–Bengali models, 2M pairs+fragments, which contain a total of 3.84M sentence pairs. Scores in bold are the highest for that translation direction.

4.3 Results

In order to evaluate if our methodology works to increase translation quality of an NMT system, we train new models using the same hyperparameters as before and evaluate them in terms of BLEU score, on the same test set as before. Table 3 shows how using our method gives us the highest BLEU score for en→bn, which is the translation direction we used to decide what data we should process for sub-sentence selection. This indicates that the added segment pairs add more value than if the same number of unchanged sentence pairs would have been added to the training data. By processing the dataset using our methodology, we reduce the training time by 65% while raising the BLEU score by 1.6. A statistical significance test performed by using MultEval (Clark et al., 2011) to do bootstrap resampling shows that our improved system, trained on less data, is significantly better than the baseline, with $p < 0.01$. It is also noteworthy that our system is only 0.6 BLEU below that of IndicTrans, reported in Section 4.1, which is almost seven times larger in terms of parameters and trained on the whole Samanantar dataset. We also tested for statistical significance between our system and IndicTrans and found that there is no statistically significant difference between the systems with $p > 0.01$. While we achieve the highest score for this translation direction using our refinement approach, the score is only slightly higher, 0.2 BLEU, than the highest score for selected subsets of the Samanantar corpus, the $3M$ sentence pair selection, and the difference is not statistically significant.

The highest scoring dataset for bn→en, significantly higher than the one created using our refinement approach, is the one containing $5M$ sentence pairs for training. That indicates that selecting data using different thresholds (see Section 4.2) for different translation directions could be beneficial before training a new MT model on the Samanantar corpus.

5 Data discarded during the Compilation of an English–Icelandic Parallel Corpus

We process the discarded English–Icelandic data in a slightly different manner. In our English–Bengali experiment we only considered subsentences within previously aligned pairs, and compared all the different concatenations of English chunks to the different combinations of Bengali chunks to find if we could raise the semantic score for the pair by removing parts from either or both sentences. For English–Icelandic, we instead consider our discarded sentences to be a comparable corpus and mine for sentence pairs from the pool of all segments in both languages. We use the approach of Steingrímsson et al. (2021b) using CLIR to create a candidate list and a logistic regression classifier to select the best sentence pairs from the list.

We start by deduplicating the discarded sentences and removing sentences that have less than three tokens that only contain alphabetical characters. This lowers the number of sentences we have to work with to $234,835$ English sentences and $242,456$ Icelandic sentences, as shown in Table 4. Next, we split all sentences into segments as we did with the English–Bengali data. As before, for splitting we use conjunctions, ‘and’ and ‘or’ for English and ‘og’ and ‘eða’ for Icelandic, as well as punctuation, the same symbols for both languages: .,;:?!()-”’|. We combine the segments into larger sentence parts and create all possible combinations of adjoining segments, ranging from single segments and up to recreating the original sentence, provided the

	English	Icelandic
Without alignments	482,975	563,381
Discarded in filtering	350,964	364,267
1. Total discarded	833,939	927,648
2. Min. three words + Deduplication	234,835	242,456
3. After sentence splits	2,793,254	2,279,111

Table 4: Number of discarded sentences used in the experiment and the resulting number of sentence segments, which are candidates for new alignments. The sentences are from the EEA subcorpus of ParIce, as described in Section 3.1.

combinations has a minimum length of three words, maximum length of 120 words, and that 70% of the tokens only contains alphabetical letters. This result in 2,793,254 unique Icelandic sentences and sentence parts and 2,279,111 English ones.

5.1 Mining for segment pairs

We start by extracting parallel sentence candidates using an inverted index-based CLIR tool called FaDA (Lohar et al., 2016), which can be applied to documents in any two languages, provided a bilingual dictionary is available. We use a publicly available English–Icelandic/Icelandic–English lexicon of $233K$ pairs (Steingrímsson et al., 2021). FaDA generates a list of 10 most likely candidates for each Icelandic and English sentence. We take an intersection of the two generated sets, resulting in 2,777,429 pairs to be inspected further. For this result, we apply the following steps:

- We remove all segment pairs with major overlap, in which more than 60% of the tokens in either language are also present in the other.
- We calculate LaBSE score for all pairs. A manual inspection of higher scoring pairs for this language pair, indicates that there may be occasional valid pairs with scores as low as 0.3, so we use that as a cutoff point.
- If two sentence pairs are identical, apart from symbols and numbers, we select the one having the higher LaBSE score.
- We calculate LASER (Schwenk, 2018), NMTScore (Vamvas and Sennrich, 2022) and WAScore for the sentences and classify them using a logistic regression classifier trained on the training set introduced in Steingrímsson et al. (2021b). We discard all pairs rejected by the classifier.

Processing Step	No. Pairs left
FaDA	2,777,429
Acceptable Overlap	1,878,202
LaBSE minimum	542,344
Remove identical	542,240
Logistic regression filter	342,066
Multiple translations removed	91,249
Subsentence removal	55,371
Language filter	36,200

Table 5: English–Icelandic sentence pairs remaining after each step of processing pairs mined from the discarded data.

Dataset	en→is BLEU	is→en BLEU
903,692 pairs (-discarded data)	43.4	54.0
939,892 pairs (+discarded data)	***43.9***	**54.3**

Table 6: Best BLEU scores for models trained with and without the sentence paired mined from discarded data. Scores in bold are the highest scores and scores in bold and italic are significantly higher than other scores.

- We check if there is more than one pair containing each English or Icelandic sentence. If so, only the highest-scoring pair in terms of LaBSE is selected.
- For each sentence pair A, we check for other sentence pairs where the sentences are subsentences of A, such that the subsentence is between 67% and 100% of the length of the original one. If we find another sentence pair, B, having an Icelandic sentence B_{is} that is a substring of A_{is} and an English sentence B_{en} which is a substring of A_{en}, we select the pair that has a higher LaBSE score and discard the other one. This way, we remove nearly identical sentence pairs originating from the same sentences.
- Finally, we run our pairs through a *fasttext* (Joulin et al., 2017) language filter, accepting pairs if the language of each sentences is correctly predicted in the top two predictions of the filter. We selected the top two predictions as we noticed that for Icelandic sentences, Icelandic was often not the first prediction, but most often in the top two predictions, unless they were somehow defective.

Table 5 shows the number of sentence pairs remaining after each processing step. After the final step, $36,200$ sentence pairs remain, mined from the $234,835$ English sentences and $242,456$ Icelandic sentences that had been previously discarded. We add these pairs to the training data previously acquired by sentence alignment and filtering, resulting in a total of $939,892$ sentence pairs. We train Transformer$_{BASE}$ models and evaluate on an in-domain evaluation set as detailed in Section 3.2. We compare the results to systems trained without the supplemental sentence pairs mined from discarded data. The systems trained with the segment pairs mined from the discarded data have slightly higher BLEU scores, but only en→is scores significantly higher than the system trained without the supplemental segment pairs. Results are given in Table 6.

6 Conclusions and Future Work

In this paper, we set out to answer whether deficient sentence pairs in a parallel corpus could be identified and refined and whether data commonly discarded when compiling parallel corpora or training NMT systems could be mined for parallel sentence pairs, that are still beneficial for training. We conducted two experiments to answer these questions. First, we tried re-evaluating sentence pairs in an English–Bengali parallel corpus in an attempt to remove extraneous data from partially parallel pairs. By partially parallel pairs we mean that a part of either sentence can align perfectly with either the whole or a part of the other sentence. Second, we collected all sentences discarded when an English–Icelandic parallel corpus was compiled, segmented them to create multiple sub-sentential variants, and treated as comparable corpora for mining parallel pairs

By using our approaches, the quality of our training corpus improved, leading to significantly better quality MT models, as measured by BLEU, when translating from English and into either Bengali or Icelandic. However, when translating into English we did not see this effect as

clearly. In the English–Bengali experiment the data selection aimed at increasing English→Bengali translation, which may explain the effect we see there, and for English→Icelandic the score rose slightly, but not significantly when the sentence pairs mined from discarded data was added. In that case the low improvement is most likely explained be the small size of the additional data, which only increased the size of the training data by 4%.

In future work we want to experiment with other methods of segmenting sentences, such as by using constituency parsing. The approach we used in this paper for segmenting was simple and easy to implement. More sophisticated segmentation may allow for more precise recombinations of sentence parts, for example by skipping parenthetical clauses or other insertions which may not be represented in both sentences. We also want to investigate whether our approaches also show positive results for other language pairs

Our experiments indicate that there is a potential in taking a second look at data that would usually be discarded, as well as in refining partially aligned sentence pairs. We showed that parallel sub-sentences are useful to acquire translation knowledge and extracting them can lead to significant improvement in performance, even using simple approaches. The methodology can thus have an impact on training future MT systems.

Finally, the training time for the different models, shown in Tables 1 and 3, indicates that smaller and more accurate training corpora have the added benefit of helping with faster convergence. In our case, training time is reduced by 65% from using the whole dataset to using our selected subset for training en→bn. The model also comes close to reaching the quality of the much larger IndicTrans model. This can translate into less need for storage and less resources at training and inference time, which is in line with a call to greener and more sustainable models of AI which consume less electricity, output fewer emissions, and perform on the whole as well as larger models, see e.g. Yusuf et al. (2021) and Jooste et al. (2022).

Acknowledgements

This work was supported by the Icelandic Centre for Research (RANNIS), grant number 228654-051, and by the ADAPT Centre for Digital Content Technology which is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online.

Artetxe, M. and Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

Bal, S., Mahanta, S., Mandal, L., and Parekh, R. (2019). Bilingual Machine Translation: English to Bengali. In Chakraborty, M., Chakrabarti, S., Balas, V. E., and Mandal, J. K., editors, *Proceedings of International Ethical Hacking Conference 2018*, pages 247–259, Singapore.

Barkarson, S. and Steingrímsson, S. (2019). Compiling and Filtering ParIce: An English-

Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland.

Barkarson, S., Steingrímsson, S., Ingimundarson, F. Á., Hafsteinsdóttir, H., and Magnússon, Á. D. (2021). ParIce dev/test sets 21.10. CLARIN-IS.

Brandt, M. D., Loftsson, H., Sigurþórsson, H., and Tyers, F. M. (2011). Apertium-IceNLP: A rule-based Icelandic to English machine translation system. In *Proceedings of the 15th Annual conference of the European Association for Machine Translation*, Leuven, Belgium.

Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, Oregon, USA.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland.

Forcada, M., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. (2011). Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and 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.

Hangya, V. and Fraser, A. (2019). Unsupervised Parallel Sentence Extraction with Parallel Segment Detection Helps Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1224–1234, Florence, Italy.

Jónsson, H. P., Símonarson, H. B., Snæbjarnarson, V., Steingrímsson, S., and Loftsson, H. (2020). Experimenting with Different Machine Translation Models in Medium-Resource Settings. In Sojka, P., Kopecek, I., Pala, K., and Horák, A., editors, *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 95–103.

Jooste, W., Haque, R., and Way, A. (2022). Knowledge Distillation: A Method for Making Neural Machine Translation More Efficient. *Information*, 13(2).

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain.

Karimi, A., Ansari, E., and Sadeghi Bigham, B. (2018). Extracting an English-Persian Parallel Corpus from Comparable Corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3477–3482, Miyazaki, Japan.

Koehn, P., Khayrallah, H., Heafield, K., and Forcada, M. L. (2018). Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Lohar, P., Ganguly, D., Afli, H., Way, A., and Jones, G. J. F. (2016). FaDA: Fast Document Aligner using Word Embedding. *The Prague Bulletin of Mathematical Linguistics*, 106:169–179.

Munteanu, D. S. and Marcu, D. (2006). Extracting Parallel Sub-Sentential Fragments from Non-Parallel Corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota.

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, Philadelphia, Pennsylvania, USA.

Paul, S. and Purkhyastha, B. S. (2020). English to Bengali Neural Machine Translation System for the Aviation Domain. *INFOCOMP Journal of Computer Science*, 19(2):78–97.

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, Belgium, Brussels.

Ramesh, G., Doddapaneni, S., Bheemaraj, A., Jobanputra, M., AK, R., Sharma, A., Sahoo, S., Diddee, H., J, M., Kakwani, D., Kumar, N., Pradeep, A., Nagaraj, S., Deepak, K., Raghavan, V., Kunchukuttan, A., Kumar, P., and Khapra, M. S. (2022). Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Schwenk, H. (2018). Filtering and Mining Parallel Data in a Joint Multilingual Space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany.

Sennrich, R. and Zhang, B. (2019). Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy.

Siddique, S., , Ahmed, T., Talukder, M. R. A., and Uddin, M. M. (2020). English to Bangla Machine Translation Using Recurrent Neural Network. *International Journal of Future Computer and Communication*, pages 46–51.

Steingrímsson, S., Loftsson, H., and Way, A. (2021a). CombAlign: a tool for obtaining high-quality word alignments. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 64–73, Reykjavik, Iceland (Online).

Steingrímsson, S., Loftsson, H., and Way, A. (2023). Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands.

Steingrímsson, S., Lohar, P., Loftsson, H., and Way, A. (2021b). Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17, Online (Virtual Mode).

Steingrímsson, S., Obrien, L. J., Ingimundarson, F. Á., Magnússon, Á. D., Andrésdóttir, Þ. D., and Eiríksdóttir, I. G. (2021). English-Icelandic/Icelandic-English glossary 21.09. CLARIN-IS.

Vamvas, J. and Sennrich, R. (2022). NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, Abu Dhabi, United Arab Emirates.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30.

Yusuf, M., Surana, P., Gupta, G., and Ramesh, K. (2021). Curb Your Carbon Emissions: Benchmarking Carbon Emissions in Machine Translation. *ArXiv*, abs/2109.12584.