Not Low-Resource Anymore: Aligner Ensembling, Batch Filtering, and New Datasets for Bengali-English Machine Translation

Tahmid Hasan*, **Abhik Bhattacharjee***, **Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman and Rifat Shahriyar**
Bangladesh University of Engineering and Technology (BUET)
{tahmidhasan, madhusudan, msrahman, rifat}@cse.buet.ac.bd,
{abhik, samin, masum}@ra.cse.buet.ac.bd

Abstract

Despite being the seventh most widely spoken language in the world, Bengali has received much less attention in machine translation literature due to being low in resources. Most publicly available parallel corpora for Bengali are not large enough; and have rather poor quality, mostly because of incorrect sentence alignments resulting from erroneous sentence segmentation, and also because of a high volume of noise present in them. In this work, we build a customized sentence segmenter for Bengali and propose two novel methods for parallel corpus creation on low-resource setups: aligner ensembling and batch filtering. With the segmenter and the two methods combined, we compile a high-quality Bengali-English parallel corpus comprising of 2.75 million sentence pairs, more than 2 million of which were not available before. Training on neural models, we achieve an improvement of more than 9 BLEU score over previous approaches to Bengali-English machine translation. We also evaluate on a new test set of 1000 pairs made with extensive quality control. We release the segmenter, parallel corpus, and the evaluation set, thus elevating Bengali from its low-resource status. To the best of our knowledge, this is the first ever large scale study on Bengali-English machine translation. We believe our study will pave the way for future research on Bengali-English machine translation as well as other low-resource languages. Our data and code are available at https://github.com/csebuetnlp/banglanmt.

1 Introduction

Recent advances in deep learning (Bahdanau et al., 2015; Wu et al., 2016; Vaswani et al., 2017) have aided in the development of neural machine translation (NMT) models to achieve state-of-the-art results in several language pairs. But a large number of high-quality sentence pairs must be fed into these models to train them effectively (Koehn and Knowles, 2017); and in fact lack of such a corpus affects the performance thereof severely. Although there have been efforts to improve machine translation in low-resource contexts, particularly using, for example, comparable corpora (Irvine and Callison-Burch, 2013), small parallel corpora (Gu et al., 2018) or zero-shot multilingual translation (Johnson et al., 2017), such languages are yet to achieve noteworthy results (Koehn et al., 2019) compared to high-resource ones. Unfortunately, Bengali, the seventh (fifth) most widely spoken language in the world by the number of (native[1]) speakers,[2] has still remained a low-resource language. As of now, only a few parallel corpora for Bengali language are publicly available (Tiedemann, 2012) and those too suffer from poor sentence segmentation, resulting in poor alignments. They also contain much noise, which, in turn, hurts translation quality (Khayrallah and Koehn, 2018). No previous work on Bengali-English machine translation addresses any of these issues.

With the above backdrop, in this work, we develop a customized sentence segmenter for Bengali language while keeping uniformity with the English side segmentation. We experimentally show that better sentence segmentation that maintains homogeneity on both sides results in better alignments. We further empirically show that the choice of sentence aligner plays a significant role in the quantity of parallel sentences extracted from document pairs. In particular, we study three aligners and show that combining their results, which we name 'Aligner Ensembling', increases recall. We introduce 'Batch Filtering', a fast and effective method for filtering out incorrect alignments. Using our new segmenter, aligner ensemble, and batch filter, we collect a total of 2.75 million high-quality parallel sentences from a wide variety of domains,

*These authors contributed equally to this work.

[1] https://w.wiki/Psq
[2] https://w.wiki/Pss

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2612–2623,
November 16–20, 2020.

more than 2 million of which were not previously available. Training our corpus on NMT models, we outperform previous approaches to Bengali-English machine translation by more than 9 BLEU (Papineni et al., 2002) points and also show competitive performance with automatic translators. We also prepare a new test corpus containing 1000 pairs made with extensive manual and automated quality checks. Furthermore, we perform an ablation study to validate the soundness of our design choices.

We release all our tools, datasets, and models for public use. To the best of our knowledge, this is the first ever large scale study on machine translation for Bengali-English pair. We believe that the insights brought to light through our work may give new life to Bengali-English MT that suffered so far for being low in resources. We also believe that our findings will also help design more efficient methods for other low-resource languages.

2 Sentence Segmentation

Proper sentence segmentation is an essential prerequisite for sentence aligners to produce coherent alignments. However, segmenting a text into sentences is not a trivial task, since the end-of-sentence punctuation marks are ambiguous. For example, in English, the end-of-sentence period, abbreviations, ellipsis, decimal point, etc. use the same symbol (.). Since either side of a document pair can contain Bengali/English/foreign text, we need a sentence segmenter to produce consistent segmentation in a language-independent manner.

Input:

কাজী মুহম্মদ ওয়াজেদের একমাত্র পুত্র ছিলেন এ. কে. ফজলুল হক।

Output:

1. কাজী মুহম্মদ ওয়াজেদের একমাত্র পুত্র ছিলেন এ.
2. কে.
3. ফজলুল হক।

Figure 1: Erroneous sentence segmentation by Polyglot

Available libraries supporting both Bengali and English segmentation, e.g., Polyglot (Al-Rfou' et al., 2013), do not work particularly well for Bengali sentences with abbreviations, which is common in many domains. For instance, Polyglot inaccurately splits the input sentence in Figure 1 into three segments, whereas the English side can successfully detect the non-breaking tokens. Not only does this corrupt the first alignment, but also causes the two broken pieces to be aligned with other sentences, creating a chain of incorrect alignments.

SegTok,[3] a rule-based segmentation library, does an excellent job of segmenting English texts. SegTok uses regular expressions to handle many complex cases, e.g., technical texts, URLs, abbreviations. We extended SegTok's code to have the same functionality for Bengali texts by adding new rules (e.g., quotations, parentheses, bullet points) and abbreviations identified through analyzing both Bengali and English side of our corpus, side-by-side enhancing SegTok's English segmentation correctness as well. Our segmenter can now address the issues like the example mentioned and provide consistent outputs in a language-agnostic manner.

We compared the performance of our segmenter on different aligners against Polyglot. We found that despite the number of aligned pairs decreased by 1.37%, the total number of words on both sides increased by 5.39%, making the resulting parallel corpus richer in content than before. This also bolsters our hypothesis that Polyglot creates unnecessary sentence fragmentation.

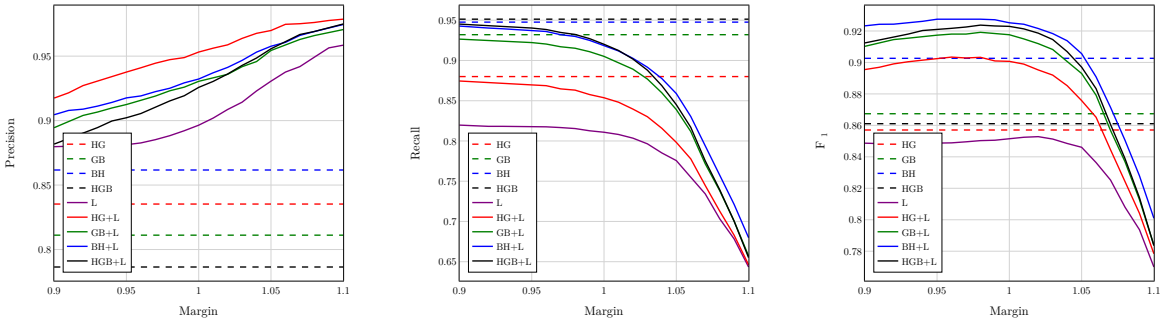
3 Aligner Selection and Ensembling

3.1 Aligner Descriptions

Most available resources for building parallel corpora come in the form of parallel documents which are exact or near-exact translations of one another. Sentence aligners are used to extract parallel sentences from them, which are then used as training examples for MT models. Abdul-Rauf et al. (2012) conducted a comparative evaluation of five aligners and showed that the choice of aligner had considerable performance gain by the models trained on the resultant bitexts. They identified three aligners with superior performance: Hunalign (Varga et al., 2005), Gargantua (Braune and Fraser, 2010), and Bleualign (Sennrich and Volk, 2010).

However, their results showed performance only in terms of BLEU score, with no indication of any explicit comparison metric between the aligners (e.g., precision, recall). As such, to make an intrinsic evaluation, we sampled 50 documents from four of our sources (detailed in section 4.2) with their sentence counts on either side ranging from 20 to 150. We aligned sentences from these documents manually (i.e., the gold alignment) and removed duplicates, which resulted in 3,383 unique sentence pairs. We then aligned the documents again with

[3] https://github.com/fnl/segtok



(a) Precision vs. Margin (b) Recall vs. Margin (c) F_1 Score vs. Margin

Figure 2: Performance metrics of ensembles with filtering

the three aligners using our custom segmenter. Table 1 shows performance metrics of the aligners.

3.2 Aligner Ensembling and Filtering

From the results in Table 1, it might seem that Hunalign should be the most ideal aligner choice. But upon closer inspection, we found that each aligner was able to correctly align some pairs that the other two had failed to do. Since we had started from a low-resource setup, it would be in our best interest if we could combine the data extracted by all aligners. As such, we 'ensembled' the results of the aligners as follows. For each combination of the aligners (4 combinations in total; see Table 2), we took the union of sentence pairs extracted by each constituent aligner of the said combination for each document. The performance of the aligner ensembles is shown in Table 2. We concatenated the first letters of the constituent aligners to name each ensemble (e.g., HGB refers to the combination of all three of them).

Table 2 shows that BH achieved the best F_1 score among all ensembles, even 0.89% above the best single aligner Hunalign. Ensembling increased the recall of BH by 8.94% compared to Hunalign, but also hurt precision severely (by 7.05%), due to the accumulation of incorrect alignments made by each constituent aligner. To mitigate this effect, we used the LASER[4] toolkit to filter out incorrect alignments. LASER, a cross-lingual sentence representation model, uses similarity scores between the embeddings of candidate sentences to perform as both aligner (Schwenk et al., 2019) and filter (Chaudhary et al., 2019). We used LASER as a

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

Aligner	Precision	Recall	F_1
Hunalign	**93.21**	85.82	**89.37**
Gargantua	84.79	69.32	76.28
Bleualign	89.41	**87.35**	88.37

Table 1: Performance metrics of aligners

Ensemble	Precision	Recall	F_1
HG	83.52	88.00	85.70
GB	81.11	93.20	86.73
BH	**86.16**	94.76	**90.26**
HGB	78.64	**95.13**	86.10

Table 2: Performance metrics of ensembles

Ensemble	Precision	Recall	F_1
L(1.02)	90.86	80.34	85.28
HG+L(0.96)	**94.09**	86.86	90.33
GB+L(0.98)	92.31	91.52	91.91
BH+L(0.96)	91.91	**93.60**	**92.75**
HGB+L(0.98)	91.52	93.23	92.37

Table 3: Performance metrics of filtered ensembles

filter on top of the ensembles, varied the similarity margin (Artetxe and Schwenk, 2019) between 0.90 to 1.10 with 0.01 increment, and plotted the performance metrics in Figure 2. We also reported the performance of LASER as a standalone aligner (referred to as L in the figure; +L indicates the application of LASER as a filter). The dashed lines indicate ensemble performance without the filter.

As Figure 2a indicates, ensembles achieve significant gain on precision with the addition of the LASER filter. While recall (Figure 2b) doesn't face a significant decline at first, it starts to take a deep plunge when margin exceeds 1.00. We balanced

between the two by considering the F_1 score (Figure 2c). Table 3 shows the performance metrics of LASER and all filtered ensembles for which their respective F_1 score is maximized.

Table 3 shows that despite being a good filter, LASER as an aligner does not show considerable performance compared to filtered ensembles. The best F_1 score is achieved by the BH ensemble with its margin set to 0.96. Its precision increased by 5.75% while trailing a mere 1.16% in recall behind its non-filtered counterpart. Compared to single Hunalign, its recall had a 7.78% gain, while lagging in precision by only 1.30%, with an overall F_1 score increase of 3.38%. Thus, in all future experiments, we used BH+L(0.96) as our default aligner with the mentioned filter margin.

4 Training Data and Batch Filtering

We categorize our training data into two sections: (1) Sentence-aligned corpora and (2) Document-aligned corpora.

4.1 Sentence-aligned Corpora

We used the corpora mentioned below which are aligned by sentences:

Open Subtitles 2018 corpus (Lison et al., 2018) from OPUS[5] (Tiedemann, 2012)

TED corpus (Cettolo et al., 2012)

SUPara corpus (Mumin et al., 2012)

Tatoeba corpus from tatoeba.org

Tanzil corpus from the Tanzil project[6]

AMARA corpus (Abdelali et al., 2014)

SIPC corpus (Post et al., 2012)

Glosbe[7] online dictionary example sentences

MediaWiki Content Translations[8]

Gnome, KDE, Ubuntu localization files

Dictionary entries from bdword.com

Miscellaneous examples from english-bangla.com and onubadokderadda.com

[5] opus.nlpl.eu
[6] tanzil.net/docs/tanzil_project
[7] https://glosbe.com/
[8] https://w.wiki/RZn

4.2 Document-aligned Corpora

The corpora below have document-level links from where we sentence-aligned them:

Globalvoices: Global Voices[9] publishes and translates articles on trending issues and stories from press, social media, blogs in more than 50 languages. Although OPUS provides sentence-aligned corpus from Global Voices, we re-extracted sentences using our segmenter and filtered ensemble, resulting in a larger amount of pairs compared to OPUS.

JW: Agić and Vulić (2019) introduced JW300, a parallel corpus of over 300 languages crawled from jw.org, which also includes Bengali-English. They used Polyglot (Al-Rfou' et al., 2013) for sentence segmentation and Yasa (Lamraoui and Langlais, 2013) for sentence alignment. We randomly sampled 100 sentences from their Bengali-English corpus and found only 23 alignments to be correct. So we crawled the website using their provided instructions and aligned using our segmenter and filtered ensemble. This yielded more than twice the data than theirs.

Banglapedia: "Banglapedia: the National Encyclopedia of Bangladesh" is the first Bangladeshi encyclopedia. Its online version[10] contains over 5,700 articles in both Bengali and English. We crawled the website to extract the article pairs and aligned sentences with our segmenter and filtered ensemble.

Bengali Translation of Books: We collected translations of more than 100 books available on the Internet with their genres ranging from classic literature to motivational speeches and aligned them using our segmenter and filtered ensemble.

Bangladesh Law Documents: The Legislative and Parliamentary Affairs Division of Bangladesh makes all laws available on their website.[11] Some older laws are also available under the "Heidelberg Bangladesh Law Translation Project".[12] Segmenting the laws was not feasible with the aligners in section 3.1 as most lines were bullet points terminating in semicolons, and treating

[9] https://globalvoices.org/
[10] https://www.banglapedia.org/
[11] bdlaws.minlaw.gov.bd
[12] https://www.sai.uni-heidelberg.de/workgroups/bdlaw/

Source	#Pairs	#Tokens(Bn)	#Tokens(En)	#Toks/Sent(Bn)	#Toks/Sent(En)
OpenSubs	365,837	2,454,007	2,902,085	6.71	7.93
TED	15,382	173,149	195,007	11.26	12.68
SUPara	69,533	811,483	996,034	11.67	14.32
Tatoeba	9,293	50,676	57,266	5.45	6.16
Tanzil	5,908	149,933	164,426	25.38	27.83
AMARA	1,166	63,447	47,704	54.41	40.91
SIPC	19,561	240,070	311,816	12.27	15.94
Glosbe	81,699	1,531,136	1,728,394	18.74	21.16
MediaWiki	45,998	3,769,963	4,205,913	81.96	91.44
Gnome	102,078	725,297	669,659	7.11	6.56
KDE	16,992	122,265	115,908	7.20	6.82
Ubuntu	5,251	22,727	22,616	4.33	4.29
Globalvoices	235,106	4,162,896	4,713,335	17.70	20.04
JW	546,766	9,339,929	10,215,160	17.08	18.68
Banglapedia	264,043	3,695,930	4,643,818	14.00	17.59
Books	99,174	1,393,095	1,787,694	14.05	18.03
Laws	28,218	644,384	801,092	22.84	28.39
HRW	2,586	55,469	65,103	21.44	25.17
Dictionary	483,174	700,870	674,285	1.45	1.40
Wiki Sections	350,663	5,199,814	6,397,595	14.83	18.24
Miscellaneous	2,877	21,427	24,813	7.45	8.62
Total	**2,751,315**	**35,327,967**	**40,739,723**	**12.84**	**14.81**

Table 4: Summary of the training corpus.

semicolons as terminals broke down valid sentences. Thus, we made a regex-based segmenter and aligner for these documents. Since most laws were exact translations with an equal number of bullet points under each section, the deterministic aligner yielded good alignment results.

HRW: Human Rights Watch[13] investigates and reports on abuses happening in all corners of the world on their website. We crawled the Bengali-English article pairs and aligned them using our segmenter and filtered ensemble.

Wiki Sections: Wikipedia is the largest multilingual resource available on the Internet. But most article pairs are not exact or near-exact translations of one another. However, such a large source of parallel texts cannot be discarded altogether. Wikimatrix (Schwenk et al., 2019) extracted bitexts from Wikipedia for 1620 language pairs, including Bengali-English. But we found them to have issues like foreign texts, incorrect sentence segmentations and alignments etc. As such, we resorted to the original source and only aligned from sections having high similarity. We translated the Bengali articles into English using an NMT model trained on the rest of our data and compared each section of an article against the sections of its English counterpart. We used SacreBLEU (Post, 2018) score as the similarity metric and only picked sections with score above 20. We then used our filtered ensemble on the resulting matches.

4.3 Batch Filtering

LASER uses cosine similarity between candidate sentences as the similarity metric and calculates margin by normalizing over the nearest neighbors of the candidates. Schwenk et al. (2019) suggested using a global space, i.e., the complete corpus for neighbor search while aligning, albeit without any indication of what neighborhood to use for filtering. In section 3.2, we used local neighborhood on document level and found satisfactory results. So we tested it with a single aligner, Hunalign,[14] on three large document sources, namely, Globalvoices (GV), JW, and Banglapedia (BP). But the local approach took over a day to filter from about 25k document pairs, the main bottleneck being the

[13] https://www.hrw.org/

[14] The optimal margin was found to be 0.95 for Hunalign.

loading time for each document. Even with several optimizations, running time did not improve much. The global approach suffered from another issue: memory usage. The datasets were too large to be fit into GPU as a whole.[15] Thus, we shifted the neighbor search to CPU, but that again took more than a day to complete. Also, the percentage of filtered pairs was quite higher than the local neighborhood approach, raising the issue of data scarcity again. So, we sought the following middle-ground between global and local approach: for each source, we merged all alignments into a single file, shuffled all pairs, split the file into 1k size batches, and then applied LASER locally on each batch, reducing running time to less than two hours.

Source	Document	1k Batch	Global
GV	**4.05**	4.60	8.03
JW	**6.22**	7.06	13.28
BP	**13.01**	14.96	25.65

Table 5: Filtered pairs (%) for different neighborhoods

In Table 5, we show the percentage of filtered out pairs from the sources for each neighborhood choice. The global approach lost about twice the data compared to the other two. The 1k batch neighborhood achieved comparable performance with respect to the more fine-grained document-level neighborhood while improving running time more than ten-folds. Upon further inspection, we found that more than 98.5% pairs from the document-level filter were present in the batched approach. So, in subsequent experiments, we used 'Batch Filtering' as standard. In addition to the document-aligned sources, we also used batch filtering on each sentence-aligned corpus in section 4.1 to remove noise from them. Table 4 summarizes our training corpus after the filtering.

5 Evaluation Data

A major challenge for low-resource languages is the unavailability of reliable evaluation benchmarks that are publicly available. After exhaustive searching, we found two decent test sets and developed one ourselves. They are mentioned below:

SIPC: Post et al. (2012) used crowdsourcing to build a collection of parallel corpora between English and six Indian languages, including Bengali. Although they are not translated by experts and have issues for many sentences (e.g., all capital letters on English side, erroneous translations, punctuation incoherence between Bn and En side, presence of foreign texts), they provide four English translations for each Bengali sentence, making it an ideal test-bed for evaluation using multiple references. We only evaluated the performance of Bn→En for this test set.

SUPara-benchmark (Mumin et al., 2018): Despite having many spelling errors, incorrect translations, too short (less than 50 characters) and too long sentences (more than 500 characters), due to its balanced nature having sentences from a variety of domains, we used it for our evaluation.

RisingNews: Since the two test sets mentioned above suffer from many issues, we created our own test set. Risingbd,[16] an online news portal in Bangladesh, publishes professional English translations for many of their articles. We collected about 200 such article pairs and had them aligned by an expert. We had them post-edited by another expert. We then removed, through automatic filtering, pairs that had (1) less than 50 or more than 250 characters on either side, (2) more than 33% transliterations or (3) more than 50% or more than 5 OOV words (Guzmán et al., 2019). This resulted in 600 validation and 1000 test pairs; we named this test set "**RisingNews**".

6 Experiments and Results

6.1 Pre-processing

Before feeding into the training pipeline, we performed the following pre-processing sequentially:

1. We normalized punctuations and characters that have multiple unicode representations to reduce data sparsity.
2. We removed foreign strings that appear on both sides of a pair, mostly phrases from which both sides of the pair have been translated.
3. We transliterated all dangling English letters and numerals on the Bn side into Bengali, mostly constituting bullet points.
4. Finally, we removed all evaluation pairs from the training data to prevent data leakage.

[15] We used an RTX 2070 GPU with 8GB VRAM for these experiments.

[16] https://www.risingbd.com/

At this point, a discussion with respect to language classification is in order. It is a standard practice to use a language classifier (e.g., Joulin et al., 2017) to filter out foreign texts. But when we used it, it classified a large number of valid English sentences as non-English, mostly because they contained named entities transliterated from Bengali side. Fearing that this filtering would hurt translation of named entities, we left language classification out altogether. Moreover, most of our sources are bilingual and we explicitly filtered out sentences with foreign characters, so foreign texts would be minimal.

As for the test sets, we performed minimal preprocessing: we applied character and punctuation normalization; and since SIPC had some sentences that were all capital letters, we lowercased those (and those only).

6.2 Comparison with Previous Results

We compared our results with Mumin et al. (2019b), Hasan et al. (2019), and Mumin et al. (2019a). The first work used SMT, while the latter two used NMT models. All of them evaluated on the SUPara-benchmark test set. We used the OpenNMT (Klein et al., 2017) implementation of big Transformer model (Vaswani et al., 2017) with 32k vocabulary on each side learnt by Unigram Language Model with subword regularization[17] (Kudo, 2018) and tokenized using SentencePiece (Kudo and Richardson, 2018). To maintain consistency with previous results, we used lowercased BLEU (Papineni et al., 2002) as the evaluation metric. Comparisons are shown in Table 6.

Model	Bn→En	En→Bn
Mumin et al. (2019b)	17.43	15.27
Hasan et al. (2019)	19.98	–
Mumin et al. (2019a)	22.68	16.26
Ours	**32.10**	**22.02**

Table 6: Comparison (BLEU) with previous works on SUPara-benchmark test set (Hasan et al., 2019 did not provide En→Bn scores)

Evident from the scores in Table 6, we outperformed all works by more than **9** BLEU points for Bn→En. Although for En→Bn the difference in improvement (**5.5+**) is not that much striking compared to Bn→EN, it is, nevertheless, commendable on the basis of Bengali being a morphologically rich language.

6.3 Comparison with Automatic Translators

We compared our models' SacreBLEU[18] (Post, 2018) scores with Google Translate and Bing Translator, two most widely used publicly available automatic translators. Results are shown in Table 7.

Model /Translator	SUPara Bn→En	SUPara En→Bn	SIPC Bn→En
Google	29.4	11.1	41.2
Bing	24.4	10.7	37.2
Ours	**30.7**	**22.0**	**42.7**

Table 7: Comparison (SacreBLEU) with automatic translators

From Table 7 we can see that our models have superior results on all test sets when compared to Google and Bing.

6.4 Evaluation on RisingNews

We performed evaluation on our own test set, **RisingNews**. We show our models' lowercased detokenized BLEU and mixedcased SacreBLEU scores in Table 8.

Metric	Bn→En	En→Bn
BLEU	39.04	27.73
SacreBLEU	36.1	27.7

Table 8: Evaluation on **RisingNews** corpus

We put great care in creating the test set by performing extensive manual and automatic quality control, and believe it is better in quality than most available evaluation sets for Bengali-English. We also hope that our performance on this test set will act as a baseline for future works on Bengali-English MT. In Figure 3, we show some example translations from the RisingNews test set.

6.5 Comparison with Human Performance

Remember that SIPC had four reference English translations for each Bengali sentence. We used the final translation as a baseline human translation and used the other three as ground truths (the fourth reference had the best score among all permutations). To make a fair comparison, we evaluated our model's score on the same three references

[17] l=32, α=0.1

[18] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a +version.1.4.1 (numrefs.4 for SIPC)

Source:	বাংলাদেশে ডিজিটাল বই প্রকাশ অনেক কারণেই গড়ে উঠেনি, যার মধ্যে রয়েছে ই-বুক রিডারের উচ্চ মূল্য এবং চাহিদার অভাব।
Reference:	In Bangladesh, publishing of digital books has not yet picked up due to a lot of reasons such as the high price of e-book readers and lack of demand.
Prediction:	The publication of digital books in Bangladesh has not been developed for many reasons, including the high price and lack of demand of e-book readers.
Source:	রোহিঙ্গাদের তাদের নিজ বাসভূমিতে নিরাপদ ও মর্যাদাপূর্ণ প্রত্যাবাসনে বাংলাদেশের অবস্থানের প্রতি জাপান পূর্ণসমর্থন ব্যক্ত করেছে।
Reference:	Japan extended its full support to Bangladesh's call for safe and dignified return of Rohingyas to their homeland.
Prediction:	Japan has expressed full support for Bangladesh's stance on safe and dignified repatriation of Rohingyas to their homelands.
Source:	In the middle of this month, situation began to deteriorate after the security forces launched an operation in the remote hilly area.
Reference:	এ মাসের মাঝামাঝি নিরাপত্তা বাহিনী দুর্গম পাহাড়ি এলাকায় দমন অভিযান শুরুর পর থেকে পরিস্থিতির অবনতি হতে শুরু করে।
Prediction:	এ মাসের মাঝামাঝি সময়ে প্রত্যন্ত পাহাড়ি এলাকায় নিরাপত্তা বাহিনী অভিযান শুরু করলে পরিস্থিতির অবনতি ঘটে।
Source:	According to a joint research report of the World Bank and the Ministry of Environment, the rate of air pollution in Dhaka is five times more than the sustainable level.
Reference:	বিশ্বব্যাংক ও পরিবেশ মন্ত্রণালয়ের এক যৌথ গবেষণা প্রতিবেদন মতে, ঢাকায় বায়ুদূষণের মাত্রা সহনীয় পর্যায়ের চেয়ে পাঁচগুণ বেশি।
Prediction:	বিশ্বব্যাংক ও পরিবেশ মন্ত্রণালয়ের যৌথ গবেষণা প্রতিবেদন অনুযায়ী, ঢাকায় বায়ু দূষণের হার গ্রহণযোগ্য মাত্রার চেয়ে পাঁচগুণ বেশি।

Figure 3: Sample translations from the RisingNews test set

instead of four. Human SacreBLEU score was 32.6, while our model scored 38.0, about **5.5** points above human judgement.

6.6 Ablation Study of Filtered Ensembles

To validate that our choice of ensemble and filter had direct impact on translation scores, we performed an ablation study. We chose four combinations based on their F_1 scores from section 3:

1. Best aligner: **Hunalign**
2. Best aligner with filter: **Hunalign+L(0.95)**
3. Best ensemble: **BH**
4. Best ensemble with filter: **BH+L(0.96)**

To ensure apples to apples comparison, we only used data from the parallel documents, i.e., Globalvoices, JW, Banglapedia, HRW, Books, and Wiki sections. Table 9 shows SacreBLEU scores along with the number of pairs for these combinations. We used the base Transformer model.

BH+L(.96) performed better than others by a noticeable margin, and the single Hunalign performed the poorest. While only having 73% pairs compared to BH, H+L(.95) stood almost on par. Despite the superiority in data count, BH could not perform well enough due to the accumulation of incorrect alignments from its constituent aligners. A clearer picture can be visualized through Figure 4. BH+L(.96) mitigated both data shortage and incorrect alignments and formed a clear envelope over the other three, giving clear evidence that the filter and the ensemble complemented one another.

Aligner /Ensemble	**#Pairs (million)**	**SUPara Bn→En**	**SIPC Bn→En**
Hunalign	1.35	20.5	33.2
H+L(.95)	1.20	21.0	33.9
BH	**1.64**	21.0	34.0
BH+L(.96)	1.44	**22.1**	**35.7**

Table 9: SacreBLEU scores for ablation study

7 Related Works

The first initiative towards machine translation for Bengali dates back to the 90s. Sinha et al. (1995) developed ANGLABHARTI, a rule-based transla-

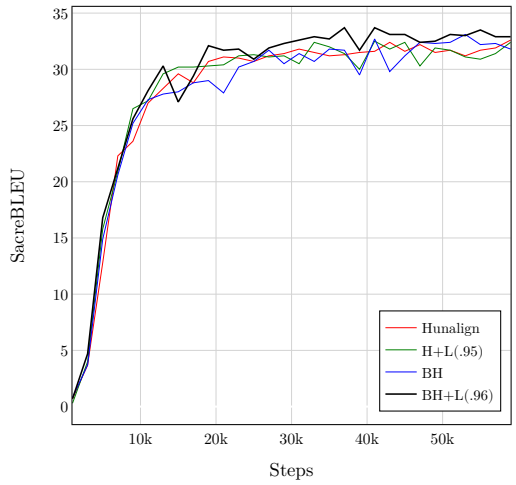


Figure 4: SacreBLEU vs Steps on SIPCdev set

tion system from English to multiple Indian languages, including Bengali. Asaduzzaman and Ali (2003); Dasgupta et al. (2004) conducted extensive syntactic analyses to write rules for constructing Bengali parse trees and designed algorithms to transfer between Bengali and English parse trees.

Subsequently, Saha and Bandyopadhyay (2005) reported an example-based machine translation approach for translating news headlines using a knowledge base. Naskar and Bandyopadhyay (2005) described a hybrid between rule-based and example-based translation approaches; here terminals would end at phrases that would then be looked up in the knowledge base.

The improved translation quality of phrase-based statistical machine translation (SMT) (Koehn et al., 2003) and the wide availability of toolkits thereof (Koehn et al., 2007) created an increased interest in SMT for Bangali-English. As SMT was more data-driven, specialized techniques were integrated to account for the low amount of parallel data for Bengali-English. Among many, Roy (2009) proposed several semi-supervised techniques; Haffari et al. (2009) used active learning to improve SMT; Islam et al. (2010) used an additional transliteration module to handle OOV words; Banerjee et al. (2018) introduced multilingual SMT for Indic languages, including Bengali.

Although NMT is currently being hailed as the state-of-the-art, very few works have been done on NMT for the Bengali-English pair. Dandapat and Lewis (2018) trained a deployable general domain NMT model for Bengali-English using sentences aligned from comparable corpora. They combated the inadequacy of training examples by data augmentation using back-translation (Sennrich et al., 2016). Hasan et al. (2019); Mumin et al. (2019a) also showed with limited parallel data available on the web that NMT provided improved translation for Bengali-English pair.

8 Conclusion and Future Works

In this work, we developed a custom sentence segmenter for Bengali, showed that aligner ensembling with batch filtering provides better performance than single sentence aligners, collected a total of 2.75 million high-quality parallel sentences for Bengali-English from multiple sources, trained NMT models that outperformed previous results, and prepared a new test set; thus elevating Bengali from its low-resource status. In future, we plan to design segmentation-agnostic aligners or aligners that can jointly segment and align sentences. We want to experiment more with the LASER toolkit: we used LASER out-of-the-box, we want to train it with our data, and modify the model architecture to improve it further. LASER fails to identify one-to-many/many-to-one sentence alignments, we want to address this. We would also like to experiment with BERT (Devlin et al., 2019) embeddings for similarity search. Furthermore, we wish to explore semi-supervised and unsupervised approaches to leverage monolingual data and explore multilingual machine translation for low-resource Indic languages.

Acknowledgements

We would like to thank the ICT Division, Government of the People's Republic of Bangladesh for funding the project and Intelligent Machines Limited for providing cloud support.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours, and Rico Sennrich. 2012. Extrinsic evaluation of sentence alignment systems. In

Proceedings of the Workshop on Creating Cross-language Resources for Disconnected Languages and Styles, pages 6–10, Istanbul, Turkey.

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

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

MM Asaduzzaman and Muhammad Masroor Ali. 2003. Morphological analysis of Bangla words for automatic machine translation. In *Proceedings of 6th International Conference on Computers and Information Technology (ICCIT)*, pages 271–276, Dhaka, Bangladesh.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, California, USA.

Tamali Banerjee, Anoop Kunchukuttan, and Pushpak Bhattacharya. 2018. Multilingual Indian language translation system at WAT 2018: Many-to-one phrase-based SMT. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

Fabienne Braune and Alexander Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING 2010)*, pages 81–89, Beijing, China. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit3: Web inventory of transcribed and translated talks. In *Proceeding of the 16th Annual Conference of the European Association for Machine Translation (EAMT 2012)*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Sandipan Dandapat and William Lewis. 2018. Training deployable general domain MT for a low resource language pair: English–Bangla. In *Proceeding of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018)*, pages 109–117, Alacant, Spain. European Association for Machine Translation.

Sajib Dasgupta, Abu Wasif, and Sharmin Azam. 2004. An optimal way of machine translation from English to Bengali. In *Proceedings of 7th International Conference on Computers and Information Technology (ICCIT)*, pages 648–653, Dhaka, Bangladesh.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, Boulder, Colorado, USA. Association for Computational Linguistics.

Md. Arid Hasan, Firoj Alam, Shammur Absar Chowdhury, and Naira Khan. 2019. Neural machine translation for the Bangla-English language pair. In *Proceedings of 22nd International Conference on Computers and Information Technology (ICCIT)*, pages 1–6, Dhaka, Bangladesh.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 262–270, Sofia, Bulgaria. Association for Computational Linguistics.

Md Zahurul Islam, Jörg Tiedemann, and Andreas Eisele. 2010. English to Bangla phrase-based machine translation. In *Proceeding of the 14th Annual Conference of the European Association for Machine Translation (EAMT 2010)*, St Raphael, France. European Association for Machine Translation.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 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. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

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.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Fethi Lamraoui and Philippe Langlais. 2013. Yet another fast, robust and open source sentence aligner. time to reconsider sentence alignment. In *Proceedings of the XIV Machine Translation Summit*, pages 77–84, Nice, France.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1742–1748, Miyazaki, Japan. European Language Resources Association (ELRA).

Md Abdullah Al Mumin, Md Hanif Seddiqui, Muhammed Zafar Iqbal, and Mohammed Jahirul Islam. 2018. SUPara-benchmark: A benchmark dataset for English-Bangla machine translation. In *IEEE Dataport*.

Md Abdullah Al Mumin, Md Hanif Seddiqui, Muhammed Zafar Iqbal, and Mohammed Jahirul Islam. 2019a. Neural machine translation for low-resource English-Bangla. *Journal of Computer Science*, 15(11):1627–1637.

Md Abdullah Al Mumin, Md Hanif Seddiqui, Muhammed Zafar Iqbal, and Mohammed Jahirul Islam. 2019b. shu-torjoma: An English ↔ Bangla statistical machine translation system. *Journal of Computer Science*, 15(7):1022–1039.

Md Abdullah Al Mumin, Abu Awal Md Shoeb, Md Reza Selim, and Muhammed Zafar Iqbal. 2012. SUPara: a balanced English-Bengali parallel corpus. *SUST Journal of Science and Technology*, 16(2):46–51.

Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2005. A phrasal EBMT system for translating English to Bengali. In *Proceedings of the Tenth Machine Translation Summit*, pages 372–279, Phuket, Thailand.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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. Association for Computational Linguistics.

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

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.

Maxim Roy. 2009. A semi-supervised approach to Bengali-English phrase-based statistical machine translation. In *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence*, page 291–294, Kelowna, Canada. Springer-Verlag.

Diganta Saha and Sivaji Bandyopadhyay. 2005. A semantics-based English-Bengali EBMT system for translating news headlines. In *Proceedings of the Tenth Machine Translation Summit*, pages 125–133, Phuket, Thailand.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, Denver, Colorado, USA.

RMK Sinha, K Sivaraman, Aditi Agrawal, Renu Jain, Rakesh Srivastava, and Ajai Jain. 1995. ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages. In *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, volume 2, pages 1609–1614.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2005)*, Borovets, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, page 6000–6010, Long Beach, California, USA.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.