The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English

Francisco Guzmán[♡♦] **Peng-Jen Chen**[♡★] **Myle Ott**[★] **Juan Pino**[♦]
Guillaume Lample[★‡] **Philipp Koehn**[■] **Vishrav Chaudhary**[♦] **Marc'Aurelio Ranzato**[★]
[♦]Facebook Applied Machine Learning [★]Facebook AI Research
[‡]Sorbonne Universités [■]Johns Hopkins University
{fguzman,pipibjc,myleott,juancarabina,guismay,vishrav,ranzato}@fb.com
phi@jhu.edu

Abstract

For machine translation, a vast majority of language pairs in the world are considered low-resource because they have little parallel data available. Besides the technical challenges of learning with limited supervision, it is difficult to *evaluate* methods trained on low-resource language pairs because of the lack of freely and publicly available benchmarks. In this work, we introduce the FLORES evaluation datasets for Nepali–English and Sinhala–English, based on sentences translated from Wikipedia. Compared to English, these are languages with very different morphology and syntax, for which little out-of-domain parallel data is available and for which relatively large amounts of monolingual data are freely available. We describe our process to collect and cross-check the quality of translations, and we report baseline performance using several learning settings: fully supervised, weakly supervised, semi-supervised, and fully unsupervised. Our experiments demonstrate that current state-of-the-art methods perform rather poorly on this benchmark, posing a challenge to the research community working on low-resource MT. Data and code to reproduce our experiments are available at https://github.com/facebookresearch/flores.

1 Introduction

Research in Machine Translation (MT) has seen significant advances in recent years thanks to improvements in modeling, and in particular neural models (Sutskever et al., 2014; Bahdanau et al., 2015; Gehring et al., 2016; Vaswani et al., 2017), as well as the availability of large parallel corpora for training (Tiedemann, 2012; Smith et al., 2013; Bojar et al., 2017). Indeed, modern neural MT systems can achieve near human-level translation performance on language pairs for which sufficient parallel training resources exist (e.g., Chinese–English translation (Hassan et al., 2018) and English–French translation (Gehring et al., 2016; Ott et al., 2018a).

Unfortunately, MT systems, and in particular neural models, perform poorly on *low-resource* language pairs, for which parallel training data is scarce (Koehn and Knowles, 2017). Improving translation performance on low-resource language pairs could be very impactful considering that these languages are spoken by a large fraction of the world population.

Technically, there are several challenges to solve in order to improve translation for low-resource languages. First, in face of the scarcity of clean parallel data, MT systems should be able to use any source of data available, namely monolingual resources, noisy comparable data, as well as parallel data in related languages. Second, we need reliable public evaluation benchmarks to track progress in translation quality.

Building evaluation sets on low-resource languages is both expensive and time-consuming because the pool of professional translators is limited, as there are few fluent bilingual speakers for these languages. Moreover, the quality of professional translations for low-resource languages is not on par with that of high-resource languages, given that the quality assurance processes for the low-resource languages are often lacking or under development. Also, it is difficult to verify the quality of the human translations as an non-native speaker, because the topics of the documents in these low-resource languages may require knowl-

[♡] Equal contribution.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6100–6113,
Hong Kong, China, November 3–7, 2019.

edge and context coming from the local culture.

In this work, we introduce new evaluation benchmarks on two very low-resource language pairs: Nepali–English and Sinhala–English. Sentences were extracted from Wikipedia articles in each language and translated by professional translators. The datasets we release to the community are composed of a tune set of 2559 and 2898 sentences, a development set of 2835 and 2766 sentences, and a test set of 2924 and 2905 sentences for Nepali–English and Sinhala–English respectively.

In §3, we describe the methodology we used to collect the data as well as to check the quality of translations. The experiments reported in §4 demonstrate that these benchmarks are very challenging for current state-of-the-art methods, yielding very low BLEU scores (Papineni et al., 2002) even using all available parallel data as well as monolingual data or Paracrawl[1] filtered data. This suggests that these languages and evaluation benchmarks can constitute a useful test-bed for developing and comparing MT systems for low-resource language pairs.

2 Related Work

There is ample literature on low-resource MT. From the modeling side, one possibility is to design methods that make more effective use of monolingual data. This is a research avenue that has seen a recent surge of interest, starting with semisupervised methods relying on back-translation (Sennrich et al., 2015), integration of a language model into the decoder (Gulcehre et al., 2017; Stahlberg et al., 2018) all the way to fully unsupervised approaches (Lample et al., 2018b; Artetxe et al., 2018), which use monolingual data both for learning good language models and for fantasizing parallel data. Another avenue of research has been to extend the traditional supervised learning setting to a *weakly supervised* one, whereby the original training set is augmented with parallel sentences mined from noisy comparable corpora like Paracrawl. In addition to the challenge of learning with limited supervision, low-resource language pairs often involve distant languages that do not share the same alphabet, or have very different morphology and syntax; accordingly, recent work has begun to explore language-independent lexical representations to improve transfer learning (Gu et al., 2018).

In terms of low-resource datasets, DARPA programs like LORELEI (Strassel and Tracey, 2016) have collected translations on several low-resource languages like English–Tagalog. Unfortunately, the data is only made available to the program's participants. More recently, the Asian Language Treebank project (Riza et al., 2016) has introduced parallel datasets for several low-resource language pairs, but these are sampled from text originating in English and thus may not generalize to text sampled from low-resource languages.

In the past, there has been work on extracting high quality translations from crowd-sourced workers using automatic methods (Zaidan and Callison-Burch, 2011; Post et al., 2012). However, crowd-sourced translations have generally lower quality than professional translations. In contrast, in this work we explore the quality checks that are required to filter *professional* translations of low-resource languages in order to build a high quality benchmark set.

In practice, there are very few publicly available datasets for low-resource language pairs, and often times, researchers *simulate* learning on low-resource languages by using a high-resource language pair like English–French, and merely limiting how much labeled data they use for training (Johnson et al., 2016; Lample et al., 2018a). While this practice enables a framework for easy comparison of different approaches, the real practical implications deriving from these methods can be unclear. For instance, low-resource languages are often distant and often times corresponding corpora are not comparable, conditions which are far from the simulation with high-resource European languages, as has been recently pointed out by Neubig and Hu (2018).

3 Methodology & Resulting Datasets

For the construction of our benchmark sets we chose to translate between Nepali and Sinhala into and out of English. Both Nepali and Sinhala are Indo-Aryan languages with a subject-object-verb (SOV) structure. Nepali is similar to Hindi in its structure, while Sinhala is characterized by extensive omissions of arguments in a sentence.

Nepali is spoken by about 20 million people if we consider only Nepal, while Sinhala is spo-

[1] https://paracrawl.eu/

ken by about 17 million people just in Sri Lanka[2]. Sinhala and Nepali have very little publicly available parallel data . For instance, most of the parallel corpora for Nepali–English originate from GNOME and Ubuntu handbooks, and account for about 500K sentence pairs.[3] For Sinhala–English, there are an additional 600K sentence pairs automatically aligned from from OpenSubtitles (Lison et al., 2018). Overall, the domains and quantity of the existing parallel data are very limited. However, both languages have a rather large amount of monolingual data publicly available (Buck et al., 2014), making them perfect candidates to track performance on unsupervised and semi-supervised tasks for Machine Translation.

3.1 Document selection

To build the evaluation sets, we selected and professionally translated sentences originating from Wikipedia articles in English, Nepali and Sinhala from a Wikipedia snapshot of early May 2018. To select sentences for translation, we first selected the top 25 documents that contain the largest number of *candidate* sentences in each source language. To this end, we defined candidate sentences[4] as: (*i*) being in the intended source language according to a language-id classifier (Bojanowski et al., 2017)[5], and (*ii*) having sentences between 50 and 150 characters. Moreover, we considered sentences and documents to be inadequate for translation when they contained large portions of untranslatable content such as lists of entities[6]. To avoid such lists we used the following rules: (*i*) for English, sentences have to start with an uppercase letter and end with a period; (*ii*) for Nepali and Sinhala, sentences should not contain symbols such as bullet points, repeated dashes, repeated periods or ASCII characters. The document set, along with the categories of documents is presented in the Appendix, Table 8.

After the document selection process, we randomly sampled 2,500 sentences for each language. From English, we translated into Nepali and Sinhala, while from Sinhala and Nepali, we only translated into English. We requested each string to be translated twice by different translators.

3.2 Quality checks

Translating domain-specialized content such as Wikipedia articles from and to low-resource languages is challenging: the pool of available translators is limited, there is limited context available to each translator when translating one string at a time, and some of the sentences can contain code-switching (e.g. text about Buddhism in Nepali or Sinhala can contain Sanskrit or Pali words). As a result, we observed large variations in the level of translation quality, which motivated us to enact a series of automatic and manual checks to filter out poor translations.

We first used automatic methods to filter out poor translations and sent them for rework. Once the reworked translations were received, we sent all translations (original or reworked) that passed the automatic checks to human quality checks. Translations which failed human checks, were disregarded. Only the translations that passed all checks were added to the evaluation benchmark, although some source sentences may have less than two translations. Below, we describe the automatic and manual quality checks that we applied to the datasets.

Automatic Filtering. The guiding principles underlying our choice of automatic filters are: (*i*) translations should be fluent (Zaidan and Callison-Burch, 2011), (*ii*) they should be sufficiently different from the source text, (*iii*) translations should be similar to each other, yet not equal; and (*iv*) translations should not be transliterations. In order to identify the vast majority of translation issues we filtered by: (*i*) applying a count-based n-gram language model trained on Wikipedia monolingual data and removing translations that have perplexity above 3000.0 (English translations only), (*ii*) removing translations that have sentence-level char-BLEU score between the two generated translations below 15 (indicating disparate translations) or above 90 (indicating suspiciously similar translations), (*iii*) removing sen-

[2]See https://www.ethnologue.com/language/npi and https://www.ethnologue.com/language/sin.

[3]Nepali has also 4K sentences translated from English Penn Tree Bank at http://www.cle.org.pk/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm, which is valuable parallel data.

[4]We first used HTML markup to split document text into paragraphs. We then used regular expressions to split on punctuation, e.g. full-stop, poorna virama (\u0964) and exclamation marks.

[5]This is a necessary step as many sentences in foreign language Wikipedias may be in English or other languages.

[6]For example, the Academy Awards page: https://en.wikipedia.org/wiki/Academy_Award_for_Best_Supporting_Actor.

tences that contain at least 33% transliterated words, (*iv*) removing translations where at least 50% of words are copied from the source sentence, and (*v*) removing translations that contain more than 50% out-of-vocabulary words or more than 5 total out-of-vocabulary words in the sentences (English translations only). For this, the vocabulary was calculated on the monolingual English Wikipedia described in Table 2.

Manual Filtering. We followed a setup similar to *direct assessment* (Graham et al., 2013). We asked three different raters to rate sentences from 0–100 according to the perceived translation quality. In our guidelines, the 0–10 range represents a translation that is completely incorrect and inaccurate, the 70–90 range represents a translation that closely preserves the semantics of the source sentence, while the 90–100 range represents a *perfect* translation. To ensure rating consistency, we rejected any evaluation set in which the range of scores among the three reviewers was above 30 points, and requested a fourth rater to break ties, by replacing the most diverging translation rating with the new one. For each translation, we took the average score over all raters and rejected translations whose scores were below 70.

To ensure that the translations were as fluent as possible, we also designed an Amazon Mechanical Turk (AMT) monolingual task to judge the *fluency* of English translations. Regardless of content preservation, translations that are not fluent in the target language should be disregarded. For this task, we then asked five independent human annotators to rate the fluency of each English translation from 1 (bad) to 5 (excellent), and retained only those above 3.Additional statistics of automatic and manual filtering stages can be found in Appendix.

3.3 Resulting Datasets

We built three evaluation sets for each language pair using the data that passed our automatic and manual quality checks: *dev* (tune), *devtest* (validation) and *test* (test). The tune set is used for hyperparameter tuning and model selection, the validation set is used to measure generalization during development, while the test set is used for the final blind evaluation.

To measure performance in both directions (e.g. Sinhala–English and English–Sinhala), we built test sets with mixed original-*translationese* (Baroni and Bernardini, 2005) on the source side. To reduce the effect of the source language on the quality of the resulting evaluation benchmark, direct and reverse translations were mixed at an approximate 50-50 ratio for the *devtest* and *test* sets. On the other hand, the *dev* set was composed of the remainder of the available translations, which were not guaranteed to be balanced. Before selection, the sentences were grouped by document, to minimize the number of documents per evaluation set.

orig lang	*dev*		*devtest*		*test*	
	uniq	tot	uniq	tot	uniq	tot
Nepali–English						
English	693	1,181	800	1,393	850	1,462
Nepali	825	1,378	800	1,442	850	1,462
	1,518	**2,559**	**1,600**	**2,835**	**1,700**	**2,924**
Sinhala–English						
English	1,123	1,913	800	1,395	850	1,465
Sinhala	565	985	800	1,371	850	1,440
	1,688	**2,898**	**1600**	**2,766**	**1700**	**2,905**

Table 1: Number of unique sentences (*uniq*) and total number of sentence pairs (*tot*) per FLORES test set grouped by their original languages.

In Table 1 we present the statistics of the resulting sets. For Sinhala–English, the *test* set is composed of 850 sentences originally in English, and 850 originally in Sinhala. We have approximately 1.7 translations per sentence. This yielded 1,465 sentence pairs originally in English, and 1,440 originally in Sinhalese, for a total of 2,905 sentences. Similarly, for Nepali–English, the *test* set is composed of 850 sentences originally in English, and 850 originally in Nepali. This yielded 1,462 sentence pairs originally in English and 1,462 originally in Nepali, for a total of 2,924 sentence pairs. The composition of the rest of the sets can be found in Table 1.

In Appendix Table 6, we present the aggregate distribution of topics per sentence for the datasets in Nepali–English and Sinhala–English, which shows a diverse representation of topics ranging from General (e.g. documents about tires, shoes and insurance), History (e.g. documents about history of the radar, the Titanic, etc.) to Law and Sports. This richness of topics increases the difficulty of the set, as it requires models that are rather domain-independent. The full list of documents and topics is also in Appendix, Table 8.

4 Experiments

In this section, we first describe the data used for training the models, we then discuss the learning settings and models considered, and finally we report the results of these baseline models on the new evaluation benchmarks.

4.1 Training Data

Small amounts of parallel data are available for Sinhala–English and Nepali–English. Statistics can be found in Table 2. This data comes from different sources. Open Subtitles and GNOME/KDE/Ubuntu come from the OPUS repository[7]. Global Voices is an updated version (2018q4) of a data set originally created for the CASMACAT project[8]. Bible translations come from the bible-corpus[9]. The Paracrawl corpus comes from the Paracrawl project[10]. The filtered version (Clean Paracrawl) was generated using the LASER model (Artetxe and Schwenk, 2018) to get the best sentence pairs having 1 million English tokens as specified in Chaudhary et al. (2019). We also contrast this filtered version with a randomly filtered version (Random Paracrawl) with the same number of English tokens. Finally, our multilingual experiments in Nepali use Hindi monolingual (about 5 million sentences) and English-Hindi parallel data (about 1.5 million parallel sentences) from the IIT Bombay corpus[11].

	Sentences	Tokens
Nepali–English		
parallel		
Bible	62K	1.5M
Global Voices	3K	75K
Penn Tree Bank	4K	88K
GNOME/KDE/Ubuntu	495K	2M
*comparable**		
Unfiltered Paracrawl	2.2M	40.6M
Clean Paracrawl	32.9K	1M
Random Paracrawl	55.3K	1M
monolingual		
Wikipedia (en)	67.8M	2.0B
Common Crawl (ne)	3.6M	103.0M
Wikipedia (ne)	92.3K	2.8M
Sinhala–English		
parallel		
Open Subtitles	601K	3.6M
GNOME/KDE/Ubuntu	46K	151K
*comparable**		
Paracrawl	3.4M	45.4M
Clean Paracrawl	47K	1M
Random Paracrawl	74.2K	1M
monolingual		
Wikipedia (en)	67.8M	2.0B
Common Crawl (si)	5.2M	110.3M
Wikipedia (si)	155.9K	4.7M

Table 2: Parallel, comparable, and monolingual data used in experiments in §4. The number of tokens for parallel and comparable corpora are reported over the English tokens. Monolingual and comparable corpora do not include any sentences from the evaluation sets. *Comparable data from Paracrawl is used only in the weakly-supervised experiments since alignments are noisy.

4.2 Training Settings

We evaluate models in four training settings. First, we consider a fully *supervised* training setting using the parallel data listed in Table 2.

Second, we consider a fully *unsupervised* setting, whereby only monolingual data on both the source and target side are used to train the model (Lample et al., 2018b).

Third, we consider a *semi-supervised* setting where we also leverage monolingual data on the target side using the standard back-translation training protocol (Sennrich et al., 2015): we train a backward MT system, which we use to translate monolingual target sentences to the source language. Then, we merge the resulting pairs of noisy (back-translated) source sentences with the original target sentences and add them as additional parallel data for training source-to-target MT system. Since monolingual data is available for both languages, we train backward MT systems in both directions and repeat the back-translation process iteratively (He et al., 2016; Lample et al., 2018a). We consider up to two back-translation iterations. At each iteration we generate back-translations using beam search, which has been shown to perform well in low-resource settings (Edunov et al., 2018); we use a beam width of 5 and individually tune the length-penalty on the *dev* set.

Finally, we consider a *weakly supervised* setting by using a baseline system to filter out Paracrawl data using LASER (Artetxe and Schwenk, 2018) by following the approach similar to Chaudhary et al. (2019), in order to augment the original training set with a possibly larger but noisier set of parallel sentences.

[7] http://opus.nlpl.eu/

[8] http://casmacat.eu/corpus/global-voices.html

[9] https://github.com/christos-c/bible-corpus/

[10] https://paracrawl.eu/

[11] http://www.cfilt.iitb.ac.in/iitb_parallel/

For Nepali only, we also consider training using Hindi data, both in a joint supervised and semi-supervised setting. For instance, at each iteration of the joint semi-supervised setting, we use models from the previous iteration to back-translate English monolingual data into both Hindi and Nepali, and from Hindi and Nepali monolingual data into English. We then concatenate actual parallel data and back-translated data of the same language pair together, and train a new model. We also consider using English-Hindi data in the unsupervised scenario. In that setting, a model is pretrained in an unsupervised way with English, Hindi and Nepali monolingual data using the unsupervised approach by Lample and Conneau (2019), and it is then jointly trained on both the Nepali–English unsupervised learning task and the Hindi-English supervised task (in both directions).

4.3 Models & Architectures

We consider both phrase-based statistical machine translation (PBSMT) and neural machine translation (NMT) systems in our experiments. All hyper-parameters have been cross-validated using the dev set. The PBSMT systems use Moses (Koehn et al., 2007), with state-of-the-art settings (5-gram language model, hierarchical lexicalized reordering model, operation sequence model) but no additional monolingual data to train the language model.

The NMT systems use the Transformer (Vaswani et al., 2017) implementation in the Fairseq toolkit (Ott et al., 2019); preliminary experiments showed these to perform better than LSTM-based NMT models. More specifically, in the supervised setting, we use a Transformer architecture with 5 encoder and 5 decoder layers, where the number of attention heads, embedding dimension and inner-layer dimension are 2, 512 and 2048, respectively. In the semi-supervised setting, where we augment our small parallel training data with millions of back-translated sentence pairs, we use a larger Transformer architecture with 6 encoder and 6 decoder layers, where the number of attention heads, embedding dimension and inner-layer dimension are 8, 512 and 4096, respectively. When we use multilingual data, the encoder is shared in the {Hindi, Nepali}–English direction, and the decoder is shared in the English–{Hindi, Nepali}direction. We regularize our models with dropout, label smoothing and weight decay, with the corresponding hyper-parameters tuned independently for each language pair. Models are optimized with Adam (Kingma and Ba, 2015) using $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 1e-8$. We use the same learning rate schedule as Ott et al. (2018b). We run experiments on between 4 and 8 Nvidia V100 GPUs with mini-batches of between 10K and 100K target tokens following Ott et al. (2018b). Code to reproduce our results can be found at https://github.com/facebookresearch/flores.

4.4 Preprocessing and Evaluation

We tokenize Nepali and Sinhala using the Indic NLP Library.[12] For the PBSMT system, we tokenize English sentences using the Moses tokenization scripts. For NMT systems, we instead use a vocabulary of 5K symbols based on a joint source and target Byte-Pair Encoding (BPE; Sennrich et al., 2015) learned using the sentencepiece library[13] over the parallel training data. We learn the joint BPE for each language pair over the raw English sentences and tokenized Nepali or Sinhala sentences. We then remove training sentence pairs with more than 250 source or target BPE tokens.

We report detokenized SacreBLEU (Post, 2018) when translating into English, and tokenized BLEU (Papineni et al., 2002) when translating from English into Nepali or Sinhala.

4.5 Results

In the supervised setting, PBSMT performed quite worse than NMT, achieving BLEU scores of 2.5, 4.4, 1.6 and 5.0 on English–Nepali, Nepali–English, English–Sinhala and Sinhala–English, respectively. Table 3 reports results using NMT in all the other learning configurations described in §4.2. There are several observations we can make.

First, these language pairs are very difficult, as even supervised NMT baselines achieve BLEU scores less than 8. Second and not surprisingly, the BLEU score is particularly low when translating into the more morphologically rich Nepali and Sinhala languages. Third, unsupervised NMT approaches seem to be ineffective on these distant language pairs, achieving BLEU scores close to 0. The reason for this failure is due to poor initialization of the word embeddings.

[12] https://github.com/anoopkunchukuttan/indic_nlp_library

[13] https://github.com/google/sentencepiece

	Supervised		Unsupervised		Semi-supervised				Weakly supervised
		+mult.		+ mult.	it. 1	it. 2	it 1. + mult.	it 2. + mult.	
English–Nepali	4.3	6.9	0.1	8.3	6.8	6.8	8.8	8.8	5.8
Nepali–English	7.6	14.2	0.5	18.8	12.7	15.1	19.8	21.5	9.6
English–Sinhala	1.2	-	0.1	-	5.2	6.5	-	-	3.1
Sinhala–English	7.2	-	0.1	-	12.0	15.1	-	-	10.9

Table 3: BLEU scores of NMT using various learning settings on *devtest* (see §3). We report detokenized SacreBLEU (Post, 2018) for {Ne,Si}→En and tokenized BLEU for En→{Ne,Si}.

Poor initialization can be attributed to to the monolingual corpora used to train word embeddings which do not have sufficient number of overlapping strings, and are not comparable (Neubig and Hu, 2018; Søgaard et al., 2018).

Fourth, the biggest improvements are brought by the semi-supervised approach using back-translation, which nearly doubles BLEU for Nepali–English from 7.6 to 15.1 (+7.5 BLEU points) and Sinhala–English from 7.2 to 15.1 (+7.9 BLEU points), and increases +2.5 BLEU points for English–Nepali and +5.3 BLEU points for English–Sinhala.

Fifth, additional parallel data in English-Hindi further improves translation quality in Nepali across all settings. For instance, in the Nepali–English supervised setting, we observe a gain of 6.5 BLEU points, while in the semi-supervised setting (where we back-translate also to and from Hindi) the gain is 6.4 BLEU points. Similarly, in the unsupervised setting, multilingual training with Hindi brings Nepali–English to 3.9 BLEU and English–Nepali to 2.5 BLEU; if however, the architecture is pretrained as prescribed by Lample and Conneau (2019), BLEU score improves to 18.8 BLEU for Nepali–English and 8.3 BLEU for English–Nepali.

Finally, the weakly supervised baseline using the additional noisy parallel data described in §4.1 improves upon the supervised baseline in all four directions. This is studied in more depth in Table 4 for Sinhala–English and Nepali–English. Without any filtering or with random filtering, BLEU score is close to 0 BLEU. Applying the a filtering method based on LASER scores (Artetxe and Schwenk, 2018) provides an improvement over using the unfiltered Paracrawl, of +5.5 BLEU points for Nepali–English and +7.3 BLEU points for Sinhala–English. Adding Paracrawl Clean to the initial parallel data improves performance by +2.0 and +3.7 BLEU points, for Nepali–English and Sinhala–English, respectively.

Corpora	**BLEU**	
	ne–en	si–en
Parallel	7.6	7.2
Unfiltered Paracrawl	0.4	0.4
Paracrawl Random	0.1	0.4
Paracrawl Clean	5.9	7.7
Parallel + Paracrawl Clean	9.6	10.9

Table 4: Weakly supervised experiments: Adding noisy parallel data from filtered Paracrawl improves translation quality in some conditions. "Parallel" refers to the data described in Table 2.

5 Discussion

In this section, we provide an analysis of the performance on the Nepali to English *devtest* set using the semi-supervised machine translation system, see Figure 1. Findings on other language directions are similar.

Fluency of references: we observe no correlation between the fluency rating of human references and the quality of translations as measured by BLEU. This suggests that the difficulty of the translation task is not related to the fluency of the references, at least at the current level of accuracy.

Document difficulty: we observe that translation quality is similar across all document ids, with a difference of 10 BLEU points between the document that is the easiest and the hardest to translate. This suggests that the random sampling procedure used to construct the dataset was adequate and that no single Wikipedia document produces much harder sentences than others.

Original vs translationese: we noticed that documents originating from Nepali are harder to translate than documents originating in English. This holds when performing the evaluation with the supervised MT system: translations of original Nepali sentences obtain 4.9 BLEU while Nepali translationese obtain 9.1 BLEU. This suggests that the existing parallel corpus is closer to English Wikipedia than Nepali Wikipedia.

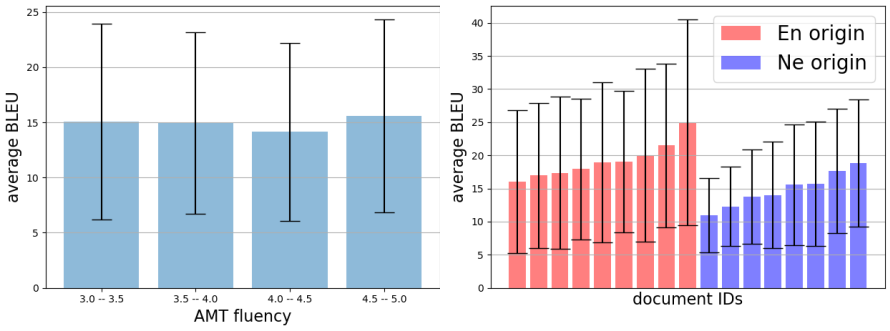

Figure 1: Analysis of the Ne→En *devtest* set using the semi-supervised machine translation system. **Left**: sentence level BLEU versus AMT fluency score of the reference sentences in English; source sentences that have received more fluent human translations are not easier to translate by machines. **Right**: average sentence level BLEU against Wikipedia document id from which the source sentence was extracted; sentences have roughly the same degree of difficulty across documents since there is no extreme difference between shortest and tallest bar. However, source sentences originating from Nepali Wikipedia (blue) are translated more poorly than those originating from English Wikipedia (red). Documents are sorted by BLEU for ease of reading.

5.1 Domain drift

To better understand the effect of domain mismatch between the parallel dataset and the Wikipedia evaluation set, we restricted the Sinhala–English training set to only the Open Subtitles portion of the parallel dataset, and we held out 1000 sentences for "in-domain" evaluation of generalization performance. Table 5 shows that translation quality on in-domain data is between 10 and 16 BLEU points higher. This may be due to both domain mismatch as well as sensitivity of the BLEU metric to sentence length. Indeed, there are on average 6 words per sentences in the Open Subtitles test set compared to 16 words per sentence in the FLORES *devtest* set. However, when we train semi-supervised models on back-translated Wikipedia data whose domain better matches the "Out-of-domain" *devtest* set, we see much larger gains in BLEU for the "Out-of-domain" set than we see on the "In-domain" set, suggesting that domain mismatch is indeed a major problem.

	Open Subtitles	**FLORES (*devtest*)**
Sinhala–English		
Supervised	23.5	7.2
Semi-sup.	28.1 (+20%)	15.1 (+210%)
English–Sinhala		
Supervised	11.0	1.2
Semi-sup.	11.8 (+7%)	6.5 (+542%)

Table 5: In-domain vs. out-of-domain translation performance (BLEU) for supervised and semi-supervised NMT models. We report BLEU on a held-out subset of 1,000 sentences from the Open Subtitles training data (see Table 2) and on *devtest* (see §3). Semi-supervised models are trained on back-translated Wikipedia data.

6 Conclusions

One of the biggest challenges in MT today is learning to translate low-resource language pairs. Research in this area not only faces formidable technical challenges, from learning with limited supervision to dealing with very distant languages, but it is also hindered by the lack of freely and publicly available evaluation benchmarks.

In this work, we introduce and freely release to the community FLORES benchmarks for Nepali–English and Sinhala–English . Nepali and Sinhala are languages with very different syntax and morphology than English; also, very little parallel data in these language pairs is publicly available. However, a good amount of monolingual data, parallel data in related languages, and Paracrawl data exist in both languages, making these two language pairs a perfect candidate for research on low-resource MT.

Our experiments show that current state-of-the-art approaches perform rather poorly on these new evaluation benchmarks, with semi-supervised and in particular multi-lingual neural methods outperforming all the other model variants and training settings we considered. We perform additional analysis to probe the quality of the datasets. We find no evidence of poor construction quality, yet observe that the low BLEU scores are partly due to the domain mismatch between the training and test datasets. We believe that these benchmarks will help the research community on low-resource MT make faster progress by enabling free access to evaluation data on actual low-resource languages and promoting fair comparison of methods.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Unsupervised statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Mikel Artetxe and Holger Schwenk. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv preprint arXiv:1812.10464*.

D. Bahdanau, K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Marco Baroni and Silvia Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmn, 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 263–268, Florence, Italy. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Conference of the Association for Computational Linguistics (ACL)*.

Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. *arXiv preprint arXiv:1611.02344*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Yun Chen, Victor OK Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. In *arXiv:1803.05567*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828.

M. Johnson, M. Schuster, Q.V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. Vigas, M. Wattenberg, G. Corrado, M. Hughes, and J. Dean. 2016. Googles multilingual neural machine translation system: Enabling zero-shot translation. In *Transactions of the Association for Computational Linguistics*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar Chris Dyer, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demo session*.

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.

G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC*. European Language Resources Association (ELRA).

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium.

Myle Ott, Michael Auli, David Granger, and Marc'Aurelio Ranzato. 2018a. Analyzing uncertainty in neural machine translation. *arXiv preprint arXiv:1803.00047*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018b. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv*, 1804.08771.

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.

Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, et al. 2016. Introduction of the asian language treebank. In *Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 2016 Conference of The Oriental Chapter of International Committee for*, pages 1–6. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Jason R Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the common crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1374–1383.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Conference of the Association for Computational Linguistics (ACL)*.

Felix Stahlberg, James Cross, and Veselin Stoyanov. 2018. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211. Association for Computational Linguistics.

Stephanie Strassel and Jennifer Tracey. 2016. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. *LREC*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Jrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229, Portland, Oregon, USA. Association for Computational Linguistics.

A List of Topics

topic	proportion (%)	
	ne–en	si–en
General	18.3	24.1
History	6.5	15.1
Science	7.4	12.7
Religion	8.9	10.5
Social Sciences	10.2	6.9
Biology	6.3	9.1
Geography	10.6	4.6
Art/Culture	6.7	8.3
Sports	5.8	6.7
Politics	8.1	N/A
People	7.4	N/A
Law	3.9	2.0

Table 6: Distribution of the topics of the sentences in the dev, devtest and test sets according to the Wikipedia document they were sampled from.

B Statistics of automatic filtering and manual filtering

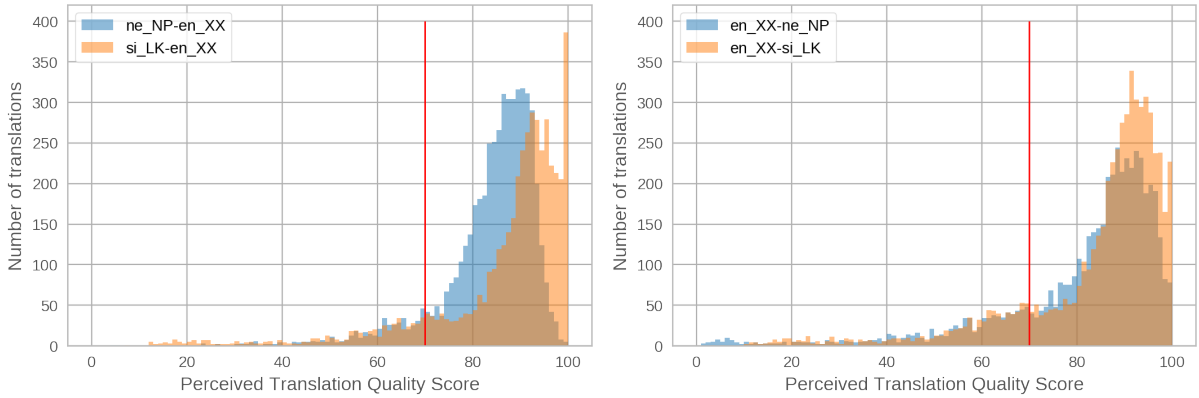

Figure 2: Histogram of averaged translation quality score. We ask three different raters to rate each sentence from 0–100 according to the perceived translation quality. In our guidelines, the 0–10 range represents a translation that is completely incorrect and inaccurate; the 11–29 range represents a translation with few correct keywords, but the overall meaning is different from the source; the 30–50 range represents a translation that contains translated fragments of the source string, with major mistakes; the 51–69 range represents a translation which is understandable and conveys the overall meaning of source string but contains typos or grammatical errors; the 70–90 range represents a translation that closely preserves the semantics of the source sentence; and the 90–100 range represents a *perfect* translation. Translations with averaged translation score less than 70 (red line) are removed from the dataset.

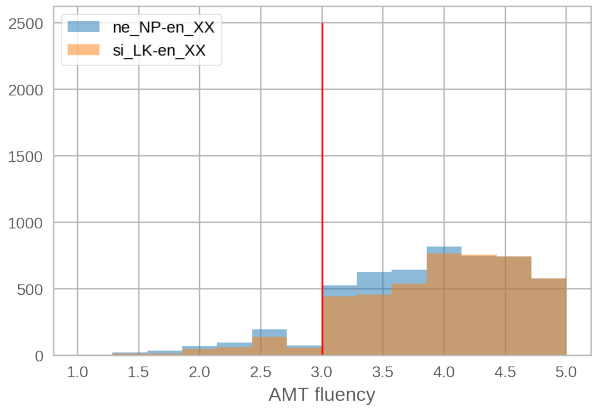

Figure 3: Histogram of averaged AMT fluency score of English translations. We ask five different raters to rate each sentence from 1–5 according to its fluency. In our guidelines, the 1–2 range represents a sentence that is not fluent, 3 is neutral, while the 4–5 range is for fluent sentences that raters can easily understand. Translations with averaged fluency score less than 3 (red line) are removed from the dataset.

	Nepali–English	English–Nepali	Sinhala–English	English–Sinhala
Automatic filtering	14%	18%	24%	7%
Manual filtering				
Translation quality	10%	19%	13%	16%
Fluency	10%	-	17%	-

Table 7: Percentage of translations that did not pass the automatic and manual filtering checks. We first use automatic methods to filter out poor translations and send those translations back for rework. We then collect translations that pass the automatic filtering and send them to two human quality checks, one for adequacy and the other for fluency. Note that the percentage of sentences that did not pass manual filtering is among those sentences that passed the automatic filtering.

C List of Wikpedia Documents

domain	document/gloss	topic
en.wikipedia.org	Astronomy	Science
en.wikipedia.org	History of radar	History
en.wikipedia.org	Shoe	General
en.wikipedia.org	Tire	General
en.wikipedia.org	Indian cuisine	Art/Culture
en.wikipedia.org	IPhone	General
en.wikipedia.org	Apollo program	History
en.wikipedia.org	Chess	General
en.wikipedia.org	Honey	General
en.wikipedia.org	Police	Law
en.wikipedia.org	Desert	Geography
en.wikipedia.org	Slavery	Social Sciences
en.wikipedia.org	Riddler	Art/Culture
en.wikipedia.org	Diving	Sports
en.wikipedia.org	Cat	Biology
en.wikipedia.org	Boxing	Sports
en.wikipedia.org	White wine	General
en.wikipedia.org	Creativity	Social Sciences
en.wikipedia.org	Capitalism	Social Sciences
en.wikipedia.org	Alaska	Geography
en.wikipedia.org	Museum	General
en.wikipedia.org	Lifeguard	General
en.wikipedia.org	Tennis	Sports
en.wikipedia.org	Writer	General
en.wikipedia.org	Anatomy	Science
si.wikipedia.org	Qoran	Religion
si.wikipedia.org	Dhammas	Religion
si.wikipedia.org	Vegetation	Science
si.wikipedia.org	Names of Colombo Students	History
si.wikipedia.org	Titanic	History
si.wikipedia.org	The Heart	Biology
si.wikipedia.org	The Ear	Biology
si.wikipedia.org	Theravada	Religion
si.wikipedia.org	WuZetian	History
si.wikipedia.org	Psychoanalisis	Science
si.wikipedia.org	Angulimala	Religion
si.wikipedia.org	Insurance	General
si.wikipedia.org	Leafart	Art/Culture
si.wikipedia.org	Communication Science	Science
si.wikipedia.org	Pharaoh Neferneferuaten	History
ne.wikipedia.org	Nelson Mandela	People
ne.wikipedia.org	Parliament of India	Politics
ne.wikipedia.org	Kailali and Kanchanpur	Geography
ne.wikipedia.org	Bhuwan Pokhari	Geography
ne.wikipedia.org	COPD	Biology
ne.wikipedia.org	KaalSarp Yoga	Religion
ne.wikipedia.org	Research Methodology in Economics	Social Sciences
ne.wikipedia.org	Essay	Social Sciences
ne.wikipedia.org	Mutation	Science
ne.wikipedia.org	Maoist Constituent Assembly	Politics
ne.wikipedia.org	Patna	Geography
ne.wikipedia.org	Federal rule systen	Law
ne.wikipedia.org	Newari Community	Art/Culture
ne.wikipedia.org	Raka's Dynasty	History
ne.wikipedia.org	Rice	Biology
ne.wikipedia.org	Breastfeeding	Biology
ne.wikipedia.org	Earthquake	Science
ne.wikipedia.org	Motiram Bhatta	People
ne.wikipedia.org	Novel Magazine	Art/Culture
ne.wikipedia.org	Vladimir Putin	Politics
ne.wikipedia.org	History of Nelali Literature	History
ne.wikipedia.org	Income tax	Law
ne.wikipedia.org	Ravi Prasjal$^+$	People
ne.wikipedia.org	Yogchudamani Upanishads$^+$	Religion
ne.wikipedia.org	Sedai$^+$	Religion

Table 8: List of documents by Wikipedia domain, their document name or English translation, and corresponding topics. The document name has an hyper-reference to the original document. $^+$ denotes a page that has been removed or no longer available at the time of this submission.

D Examples from *devtest*

En→Ne		
Source		It has automatic spell checking and correction, predictive word capabilities, and a dynamic dictionary that learns new words.
References	**A**	यसमा स्वचालित हिज्जे जाँच र सुधार छ , भविष्यवाणी शब्द क्षमताहरू , र गतिशील शब्दकोश हुन्छ जसले नयाँ शब्दहरू सिक्छ ।
	B	यसमा स्वचालित हिज्जे जाच्ने तथा सच्याउने , शब्दहरूको अनुमान गर्ने , तथा नयाँ शब्दहरू सिक्ने स्फुर्त शब्दकोश हुन्छ ।
System		यसमा स्वचालित हिज्जे जाँच र सुधार , पूर्वानुमान शब्द क्षमता र नयाँ शब्द सिक्ने गतिशील शब्दकोश छ ।
Source		The academic research tended toward the improvement of basic technologies, rather than their specific applications.
References	**A**	शैक्षिक अनुसन्धानले उनीहरूको विशिष्ट अनुप्रयोगहरूको सट्टा आधारभूत प्रविधिको सुधारको पक्षमा जोड दिए ।
	B	यो शैक्षणिक अनुसन्धान सामान्य प्रविधिको सुधार तर्फ ढलकिएको छ , नाकि तिनिहरूको विशेश प्रयोग तर्फ ।
System		प्राध्यापक अनुसन्धानले उनीहरूको विशिष्ट अनुप्रयोगमन्दा पनि आधारभूत प्रविधिको सुधारतिर टेवा पुर्यायो ।
Ne→En		
Source		पुरानो समयमा राजालाई सल्लाह दिने सभा ' संसद ' कहलाउँथ्यो ।
References	**A**	In the past, the assembly that advised the king were called 'parliament'.
	B	In old times the counsil that gave advice to the king was called 'parliament'.
System		In old times the council of counsel to the king was 'Senate'.
Source		कार्यकर्ताका रुपमा अफ्रिकन नेशनल कंग्रेसमा आबद्ध भए ।
References	**A**	As a worker African Mandela joined the Congress party.
	B	He joined the African National Congress as a activist.
System		As a worker, he joined the African National Congress.
En→Si		
Source		Iphone users can and do access the internet frequently, and in a variety of places.
References	**A**	අයිෆෝන් භාවිත කරන්නන්ට නිතරම සහ විවිධ ස්ථානවලදී අන්තර්ජාලයට පිවිසිය හැකිය .
	B	Iphone පරිශීලකයින් හට නිතරම විවිධ ස්ථානවලදී අන්තර්ජාලය වෙත පිවිසීමට හැකිය .
System		අයිෆෝන් භාවිතා කරන්නන්ට පුළුවන් වගේම අන්තර්ජාලයට නිතර පිවිසෙන්නත් පුළුවන් . ඒ වගේම විවිධ තැන්වල
Source		In Serious meets, the absolute score is somewhat meaningless.
References	**A**	සැබෑ තරඟ වලදී ලකුණු සැසඳීම තේරුමක් නැති ක්රියාවකි .
	B	වැදගත් උළෙල වල , නිරපේක්ෂ ලකුණු තරමක් නිෂ්ඵල ය .
System		සීරියස් හමුවේ , නිරපේක්ෂ ලකුණු කිසියම් තේරුමක් නැති
Si→En		
Source		තර්ජන , ශාරීරික හිංසනය , දේපල හානිය , පහර දීම සහ මරාදැමීම මෙම දඬුවම්ය .
References	**A**	Threatening, physical violence, property damage, assault and execution are these punishments.
	B	Threats, bodily violence, property damages, assaults and killing are these punishments.
System		Threats, physical harassment, property damage, strike and killing this punishment.
Source		අධ්‍ය යාපනයෙන් පසු හෝ පවුලේ යුතුකම් ඉටු කරන්නට හෝ රෝග තත්වයන් නිසා සංඝයා උපසම්පදාවෙන් නිතරම ඉවත් වෙති .
References	**A**	After education priests leave ordination in order to fulfill duties to the family or due to sickness.
	B	Sangha is often abandoned because of education or after fulfilling family responsibilities or because of illness.
System		After education or to fulfill the family's disease or disease conditions, the companion is often removed from substance.

Table 9: Examples of sentences from the En-Ne, Ne-En, En-Si and Si-En *devtest* set. System hypotheses (System) are generated using the semi-supervised model described in the main paper using beam search decoding.