Improving Machine Translation using Corpus Filtering: A Survey

Akshay Batheja, Pushpak Bhattacharyya

Department of Computer Science and Engineering
CFILT, Indian Institute of Technology Bombay
Mumbai, India
{akshaybatheja, pb}@cse.iitb.ac.in

Abstract

Web-crawled data serves as a valuable resource for training machine translation models, providing parallel corpora. However, this data is inherently noisy, and recent research has revealed the heightened sensitivity of neural machine translation systems to such noise compared to traditional statistical methods. To address this challenge, the task of Parallel Corpus Filtering (PCF) aims to extract high-quality parallel corpora from noisy pseudo-parallel corpora. In this paper, we present an extensive analysis of different approaches proposed for parallel corpus filtering. By examining previous works, we establish a roadmap that not only summarizes the existing methodologies but also lays the foundation for future research in this domain. The findings of this paper shed light on the complexities of PCF and offer valuable insights into the development of robust and accurate parallel corpus filtering techniques, thereby advancing the field of machine translation.

1 Introduction

In recent times, Neural MT has shown excellent performance, having been trained on a large amount of parallel corpora (Dabre et al., 2020). However, not all language pairs have a substantial amount of parallel data. Hence, we have to rely on the noisy web-crawled corpora for low-resource languages. Given the limited availability of clean parallel data, the use of multilingual noisy data, such as web-crawls, as an alternative for training translation systems becomes increasingly important.

Recently, there is an increased interest in the filtering of noisy parallel corpora to increase the amount of data that can be used to train translation systems (Koehn et al., 2018). The Shared Task on Parallel Corpus Filtering and Alignment at the Conference for Machine Translation (WMT 2018, WMT 2019, WMT 2020) was organized to promote research to make learning from noisy data more viable for low-resource languages.

1.1 Motivation

The Deep Neural architecture has become the most widely used architecture to build a Machine Translation (MT) model. The performance of a data-driven machine translation system is influenced by the quality and quantity of data available for training. The web-crawled data available for low-resource languages is undoubtedly high in quantity, but their quality varies a lot. This motivates us to extract high-quality parallel corpora from web-crawled pseudo-parallel sources, with the goal of improving the quality of the machine translation model in comparison to the model trained solely on noisy pseudo-parallel corpora.

2 Background and Terminology

2.1 Machine Translation

Machine Translation aims to automatically translate text from one language to text in another with the help of some software. The field of MT has experienced a significant paradigm shift in recent years. The developments in the field of MT have reduced the barrier of language. The fundamental paradigms of machine translation are:

1. **Rule Based Machine Translation:** Machine Translation follows the analysis-transfer-generation (ATG) (Bhattacharyya, 2015) process. In RBMT, human experts create all the rules manually and are responsible for the translation.

2. **Example Based Machine Translation:** In this approach, a parallel corpus is used. For a given input sentence, fragments of the phrases are matched with the existing parallel sentences in the corpus. Now, the translations of the matched fragments are picked up and put together to form a complete translation.

3. **Statistical Machine Translation:** In this methodology, a parallel corpus is utilized to

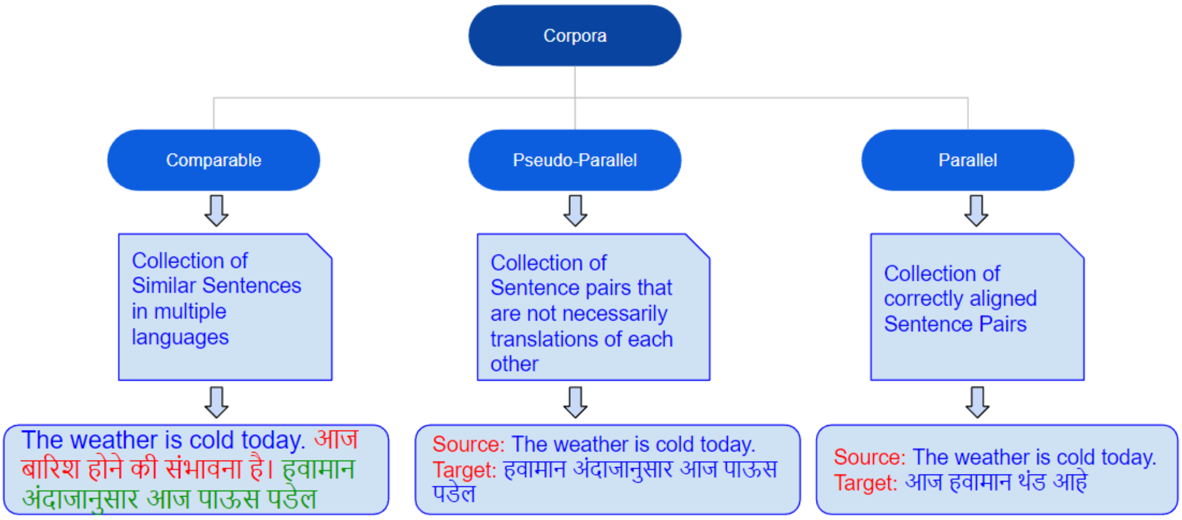

Figure 1: Types of Corpora

acquire mappings between words and phrases in both the source and target sentences, employing a probabilistic model. Statistical Machine Translation (SMT) encompasses several key elements, including a Language Model, Translation Model, Decoder, and parameter estimation. The model learns from the parallel corpus to construct a phrase table, which serves as a reference for translating input sentences based on probability values.

4. **Neural Machine Translation (NMT):** NMT aims to develop an end-to-end model using Neural Architecture to effectively translate text between different languages. To train an NMT system, a substantial amount of parallel data is required. The cutting-edge model for Machine Translation (MT) at present is built upon neural architecture.

5. **Multilingual Neural Machine Translation:** The goal of multilingual NMT is to train a single, end-to-end model that can produce translations for multiple languages.

2.2 Comparable Corpora

A comparable corpus is a collection of similar sentences in multiple languages. For instance, sentences crawled for Wikipedia's multilingual pages. Such sentences need not be exact translations of each other or aligned but they refer to the same topic in different languages. We discuss the extraction process of comparable corpora in section 10.

2.3 Parallel Corpora

Parallel Corpora is a collection of aligned sentence pairs. For instance, Hindi-Marathi parallel corpus refers to a dataset that has Hindi sentences at the source side and Marathi sentences at the target side. The sentence pairs are semantically similar and are of good quality.

2.4 Pseudo-Parallel Corpora

Pseudo-Parallel Corpora is a collection of sentence pairs that are not necessarily aligned. Thus pseudo-parallel corpora contain noisy sentence pairs that can be misaligned, disfluent and inadequate.

2.5 Parallel Corpus Filtering

The objective of Parallel Corpus Filtering (PCF) is to retrieve high-quality parallel data from pseudo-parallel corpora that contain noise. This can be performed in the following ways:

1. **Rule-based PCF:** In the rule-based approach for Parallel Corpus Filtering (PCF), we employ straightforward rules based on sentence length and linguistic features to eliminate noisy sentence pairs.

2. **Neural PCF:** In this method, we train a neural model to score the sentence pairs based on their semantic similarity.

2.6 Phrase Table Injection

In this method, we train Phrase Based Statistical Machine Translation model to generate a Phrase Table for a language pair. Then, we augment the phrase pairs retrieved from the phrase table, to the parallel corpora. This is known as Phrase Table Injection.

2.7 Quality Estimation

Quality Estimation (QE) involves assessing the quality of a translation in the absence of a reference translation. In their work, (Ranasinghe et al., 2020) introduced a QE framework based on cross-lingual transformers. This model takes both the source sentence and its translation as input and generates either a Direct Assessment score or an HTER score.

2.8 Language Agnostic Bert Sentence Embedding

LaBSE, a multilingual embedding model, provides support for 109 languages, including several Indic languages. A multilingual embedding model is a powerful approach that enables the mapping of sentences from different languages into a shared vector space.

2.9 Automatic Post-Editing

The purpose of **Automatic Post Editing** (APE) is to automatically identify and correct errors in Machine Translation (MT) outputs. Deoghare and Bhattacharyya (2022) introduced a curriculum training strategy for training the APE system.

3 Parallel Corpus Filtration techniques in SMT

The paper [(Skadiņa et al., 2012)] discusses the creation of comparable corpora and parallel data extraction from the comparable corpora. The Comparable corpora is collected from the web through Wikipedia and News Corpora.

3.1 Comparability Metric

Comparability Metric is used to evaluate the quality of Comparable Corpora. We construct feature vectors based on the lexical information and document structure. Then, we compute Cosine similarity on these feature vectors to compute the comparability scores. Now, based on the threshold value of this similarity score, the comparable corpora is ranked as either parallel or strongly comparable, or weakly comparable.

3.2 Extracting Parallel data

Parallel data is extracted in the following two ways from the comparable corpora:

1. Phrase Table Injection: Extracting parallel phrases and sentences using EMACC (Expectation-Maximization Alignment for Comparable Corpora) tool.

2. Extracting named entities and terminological units: No matter how weak comparable corpora are, they still can contain useful translational equivalences for named entities.

3.3 Experimental Result

An experiment is performed on EN-DE (English-German) domain-adapted SMT for the automotive industry domain. The parallel data extracted from comparable corpora for the automative industry domain is used for training the model. In the results 2, we see that the baseline model, which is trained without the extracted parallel data, lags behind the automotive extracted model by 7 BLEU score points.

System	BLEU
Baseline	18.81
Automotive extracted	25.44

Figure 2: Evaluation of narrow domain SMT system enriched with data from comparable corpus.

3.4 Parallel Corpus Filtration Techniques in NMT

In this section, we will look at the neural approaches for filtering parallel corpus to improve the performance of NMT systems.

4 LaBSE based Filtering

Language Agnostic BERT Sentence Embedding model [(Feng et al., 2020)] is a multilingual embedding model that supports 109 languages including some Indic languages. A multilingual embedding model is an effective method that maps sentences of various languages over the same vector space. This allows the model to leverage semantic information of multiple languages for better language understanding.
Some of the previous approaches for generating sentence embeddings are **LASER** and **m-use**. Both of the models directly map sentences from one language to another to obtain sentence embeddings. With the use of pre-training techniques MLM and TLM, the LaBSE model is trained on a huge dataset

due to which it can generate embeddings even for zero-shot languages.

Model Architecture

The architecture of this model is based on the Bi-directional dual encoder with additive margin softmax loss. We can see the architecture as shown in the figure 3.

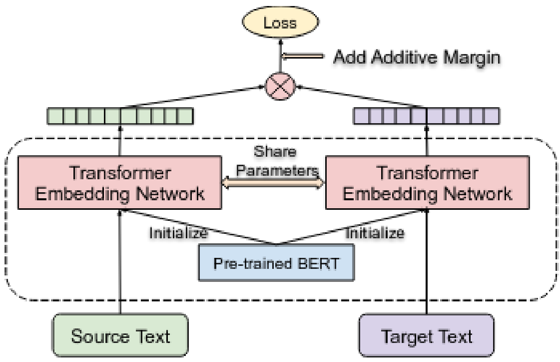

Figure 3: LaBSE Model Architecture

Training Pipeline

Firstly, the Multilingual BERT model is trained on 109 languages for MLM (Masked Language Model) task. Then the obtained BERT encoders are used in parallel at the source and target to finetune the Translation Ranking Task. So, it combines the strategies like pre-training and finetuning with bi-directional dual encoders translation ranking model.

Translation Ranking task

The goal of this task is as follows:

1. To rank all the target sentences in order of their compatibility with the score.
2. The objective is to maximize the similarity between the source sentence and its authentic translation while minimizing it with other sentences through the process of negative sampling.
3. The dual-encoder architecture encodes two sequences using parallel encoders and then utilizes a dot product to calculate the similarity score between the two encodings.
4. Bidirectional means it takes compatibility scores in both directions i.e, from source to target as well as target to source and the individual losses are summed :

$$Loss = L + L'$$

Additive Margin Softmax

It introduces a parameter $\boldsymbol{m}$ in the original softmax loss function to increase the separability between the vectors in the vector space. The loss function is given as given below. We can see that m is subtracted only from the positive sample and not from the negative samples. This is responsible for the classification boundary.

$$\begin{aligned}\mathcal{L}_{AMS} &= -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s\cdot(cos\theta_{y_i}-m)}}{e^{s\cdot(cos\theta_{y_i}-m)}+\sum_{j=1,j\neq y_i}^{c} e^{s\cdot cos\theta_j}} \\ &= -\frac{1}{n}\sum_{i=1}^{n} log\frac{e^{s\cdot(W_{y_i}^T \boldsymbol{f}_i-m)}}{e^{s\cdot(W_{y_i}^T \boldsymbol{f}_i-m)}+\sum_{j=1,j\neq y_i}^{c} e^{sW_j^T \boldsymbol{f}_i}}\end{aligned}$$

Experimental Results

Figure 4 shows the Tatoeba bitext retrieval task compared against the prior state-of-the-art bilingual models. It is evident that LaBSE surpasses other models, exhibiting a state-of-the-art average accuracy of 87.3% across all languages.

Model	14 Langs	36 Langs	82 Langs	All Langs
m~USE*	93.9	–	–	–
LASER	95.3	84.4	75.9	65.5
LaBSE	95.3	95.0	87.3	83.7

Figure 4: Average accuracy(%) on Tatoeba Datasets

5 Distilled PML based filtering

5.1 Distilled Paraphrase Multilingual Model

Distilled Paraphrase Multilingual Model is a Sentence BERT model extended to multiple languages using multilingual knowledge distillation. In the paper [(Reimers and Gurevych, 2020)], a new method is presented to generate a multilingual embedding model. Using this method we can extend the existing sentence embedding model (Monolingual/Multilingual) to new languages.
The Teacher-Student model architecture is used while training the model as shown in the fig. 5

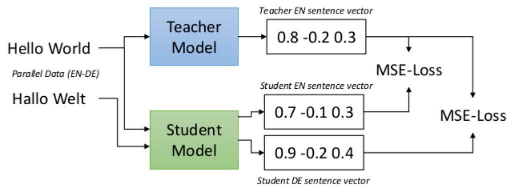

Figure 5: Teacher Student Model Architecture

Teacher-Student Model architecture

1. Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector.

2. It requires a Model M (Teacher) that maps sentences in one or more source languages to a dense vector space.

3. It also requires parallel sentences (s_n, t_n), where s_i is a sentence in source language and t_i is a sentence in one of the target languages.

4. A student model M'(t_i) and M(s_i) produces the similar sentence vector and M'(s_i) and M(s_i) produces the similar sentence vector.

5. For a Batch size B, the MSE loss is minimized as given below

$$\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}}\left[(M(s_j)-\hat{M}(s_j))^2+(M(s_j)-\hat{M}(t_j))^2\right]$$

6. During training, Sentence BERT is chosen as a Teacher model and XLM-R (XLM-RoBERTa) is chosen as a student model.

7. So, a student model is trained using XLM-R and further fine-tuned on STS(Semantic Text Similarity) and NLI (Natural Language Inference) tasks using English SBERT (Sentence BERT) model.

It was shown that this model performs better than LaBSE model on **Semantic Text Similarity (STS)** task Benchmark data while LaBSE performed better in **BUCC** (Zweigenbaum et al., 2017) bitext retrieval task.

Experimental Results

1. **Multilingual Semantic Text Similarity**

   The goal of this task is to assign a similarity score to a sentence pairs. For an instance, zero score means the sentence pairs are not related and five means they are semantically equivalent.
   This experiment is performed on STS 2017 dataset which contains annotated pairs for EN-EN, AR-AR, ES-ES, EN-AR, EN-ES, EN-TR. This dataset is further extended to EN-FR, EN-IT, and EN-NL. The Spearman rank correlation is calculated between the cosine similarities of the sentence representations generated by the model and the gold labels for STS 2017 dataset.

2. **BUCC: Bitext retrieval** This task aims to extract parallel sentences from a given comparable corpora. The dataset from BUCC bitext mining task is used to extract parallel sentences between an English corpus and other four languages. The results of this experiment are shown in figure 8.

3. **Tatoeba: Similarity Search** This task aims to extract parallel sentences for low resource languages. For evaluation, test setup from LASER is used. The dataset contains upto 1000 English-aligned sentence pairs for various languages. The evaluation is done by finding most similar sentences for all language pairs using cosine similarity. Accuracy is computed for both directions in the language pair.

6 Extracting In-Domain Parallel Corpora

The continuous increase in data through different sources like the web and news, results in larger generic models. Such generic models perform poorly in domain-specific cases. The paper (**?**), introduced an approach to select In-domain data from general-domain corpora in order to improve MT. This method ranks generic-domain sentences based on how similar they are to domain-specific monolingual corpora. Then, we choose K sentences that have the best similarity score.

Data Selection Pipeline

The In-domain Data selection Pipeline is as follows:

Model	EN-EN	ES-ES	AR-AR	Avg.
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
Knowledge Distillation				
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	82.0
XLM-R ← SBERT-paraphrases	88.8	86.3	79.6	**84.6**
Other Systems				
LASER	77.6	79.7	68.9	75.4
mUSE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4

Figure 6: Spearman rank correlation between the cosine similarity of sentence representations and the gold labels for STS 2017 dataset

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
Knowledge Distillation								
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	77.9
XLM-R ← SBERT-paraphrases	82.3	84.0	80.9	83.1	84.9	86.3	84.5	**83.7**
Other Systems								
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mUSE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5

Figure 7: Spearman rank correlation between the cosine similarity of sentence representations and the gold labels for STS 2017 dataset

1. The data selection method evaluates the similarity between general-domain sentences and in-domain monolingual data to rank the sentences accordingly.

2. This pipeline is mainly constructed of three components:

   (a) **A Contextual Sentence Embedding Component:** In this stage, we compute the sentence embeddings for the in-domain monolingual and generic-domain data for the corresponding language using **SBERT** (Sentence BERT). The embeddings generated by the SBERT model are of higher dimension. So, we bring this dimension to a smaller size by keeping only the principal components using **PCA** (Principal Component Analysis) algorithm. An illustration for this step is shown in figure 10.

   (b) **Semantic Search Component:** After generating embeddings for both monolingual and generic domain data, the Cosine Similarity Score is calculated between each in-domain sentence and each out-of-domain sentence. Using this score, the generic-domain corpora is then ranked accordingly. An illustration of this step is shown in figure 11

   (c) **Ranking In-Domain data component:** After generating similarity scores, the sentences corresponding to the top 6 scores are extracted from the out-of-domain data. These selected sentences are then referred to as in-domain sentences. An illustration for this step is

Model	DE-EN	FR-EN	RU-EN	ZH-EN	Avg.
mBERT mean	44.1	47.2	38.0	37.4	41.7
XLM-R mean	5.2	6.6	22.1	12.4	11.6
mBERT-nli-stsb	38.9	39.5	26.4	30.2	33.7
XLM-R-nli-stsb	44.0	51.0	51.5	44.0	47.6
Knowledge Distillation					
XLM-R $\leftarrow$ SBERT-nli-stsb	86.8	84.4	86.3	85.1	85.7
XLM-R $\leftarrow$ SBERT-paraphrase	90.8	87.1	88.6	87.8	88.6
Other systems					
mUSE	88.5	86.3	89.1	86.9	87.7
LASER	95.4	92.4	92.3	91.7	93.0
LaBSE	95.9	92.5	92.4	93.0	93.5

Figure 8: F_1 score on the BUCC bitext mining task

Model	KA	SW	TL	TT
LASER				
en $\rightarrow$ xx	39.7	54.4	52.6	28.0
xx $\rightarrow$ en	32.2	60.8	48.5	34.3
XLM-R $\leftarrow$ SBERT-nli-stsb				
en $\rightarrow$ xx	73.1	85.4	86.2	54.5
xx $\rightarrow$ en	71.7	86.7	84.0	52.3

Figure 9: Accuracy on the Tatoeba test set in both directions (en to target language and vice versa).

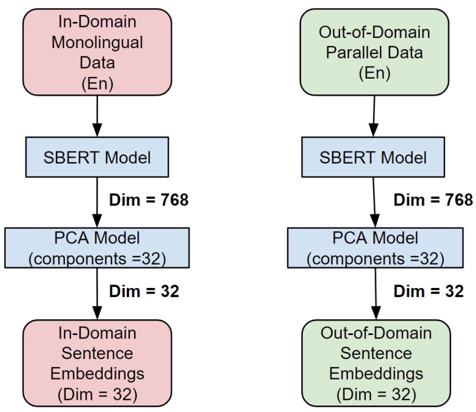

Figure 10: Context Sentence Embedding Component

shown in figure 12.

Experimental Results

The experiment was conducted with the following datasets:

1. TED training dataset (IWSLT 2014), which consists of 179K sentences. This dataset is considered the In-Domain dataset.

2. WMT training dataset, which consists of 30M sentence pairs. This dataset is considered a generic-domain (out-of-domain) dataset.

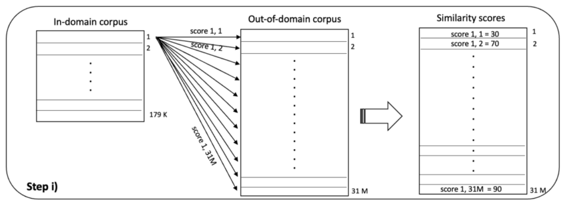

Figure 11: Semantic Search Component

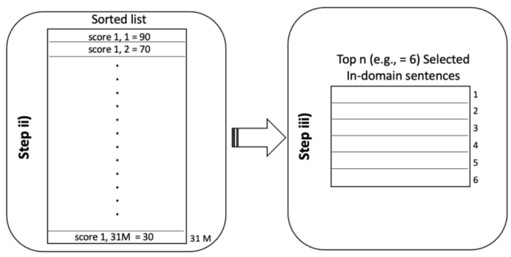

Figure 12: Ranking In-Domain data Component

We can see in the figure 13 that the NMT model trained on subcorpora ($Top6 + Top5 + Top4..$) with corpus size 1M performs comparably to the Baselines NMT Domain Adaptation models, which are trained on a relatively much larger corpus.

7 Types of Noise in a Pseudo-Parallel Corpus

Herold et al. (2022) studied various types of noise present in the Pseudo-Parallel corpora and investigated if the current filtering systems remove all types of noise.

Types of Noise

Noise can be introduced into the clean training data in the following ways:

Systems	Number of Sentences	NMT- Test Set 2010			NMT- Test Set 2011		
		BLEU↑	TER↓	CHRF2↑	BLEU↑	TER↓	CHRF2↑
S1:ID	179K	31.9	56.6	57.0	38.3	49.7	61.0
S2:OOD	31.0M	25.8	66.1	53.0	30.7	59.3	47.0
S3:ID+OOD	31.1M	26.0	62.9	54.0	30.9	56.8	58.0
B4:Luong	17.9M	32.2	N/A	N/A	35.0	N/A	N/A
B5:Axelrod	9.0M	32.2	N/A	N/A	35.5	N/A	N/A
B6:Chen	7.3M	30.3	N/A	N/A	33.8	N/A	N/A
B7:Wang	3.7-7.3M	32.8	N/A	N/A	36.5	N/A	N/A
Top1	179K	21.8	69.8	50.0	25.6	64.0	53.0
Top2+top1+...	358K	26.7	63.4	54.0	31.3	57.1	57.0
Top3+top2+...	537K	29.1	60.4	56.0	34.3	53.9	60.0
Top4+top3+...	716K	30.7	59.5	57.0	35.6	52.6	61.0
Top5+top4+...	895K	30.9	59.1	57.0	**36.7**	**51.5**	**62.0**
Top6+top5+...	1.0M	**31.3**	**58.3**	**58.0**	36.5	50.9	62.0

Figure 13: English→French: Evaluation scores for NMT system

1. **Misaligned Sentences**: Shuffle target side of the clean corpus.
2. **Misordered Words**: Shuffle words of either source or the target sentence.
3. **Wrong language**: Add sentence pairs of different languages.
4. **Untranslated**: Convert src-tgt corpus to src-src or tgt-tgt.
5. **Raw Crawled Data**: Add data from unfiltered web crawled corpus.
6. **Over/Under-translation**: Remove second half of src or tgt sentence.
7. **Synthetic Translations**: Add machine-translated sentences crawled from different websites.

Experiment Results

Two state-of-the-art experiments were conducted, namely, Cross-Entropy based Filtering and LASER based Filtering. The dataset used for the experiments are mentioned below.

- **De→En**: Dataset from WMT2017 News Translation task Randomly selected 350K sentence pairs to create the noise categories.
- **Km→En**: Dataset from WMT2020 parallel corpus filtering task Extracted 20K sentence pairs to create synthetic noisy datasets.
- **Raw Crawled data**: 20K sentence pairs from the ParaCrawl project.

The results are shown in figure 14 and 15.

8 Quality Estimation

Quality Estimation (QE) involves assessing the quality of a translation in the absence of a reference translation. In their work, (Ranasinghe et al., 2020) introduced a QE framework based on cross-lingual transformers. This model takes both the source sentence and its translation as input and generates either a Direct Assessment score or an HTER score.

Model Architecture

Two architectures are proposed in the work, namely, **MTransQuest** and **STransQuest**. XLM-Roberta model is used in both architectures. The two architectures shown in 16 are as follows:

- **MTransQuest:**
  - Using a [SEP] token, the original text and its translation are combined to form the input.

Noise Category	Corrupted Side	Filtering Accuracy				
		Cross Entropy	LASER	Language ID Filtering + none	+ CE	+ LASER
Misaligned Sentences	none	65% / 65%	72% / 76%	50%	64% / 65%	71% / 75%
Misordered Words	src	89% / 89%	62% / 70%	50%	88% / 88%	61% / 70%
	tgt	95% / 96%	62% / 70%	50%	93% / 94%	61% / 70%
Wrong Language	src	89% / 89%	51% / 54%	97%	97% / 97%	97% / 97%
	trg	87% / 87%	54% / 60%	96%	96% / 96%	96% / 96%
Untranslated	src	62% / 62%	15% / 50%	97%	97% / 97%	97% / 97%
	trg	93% / 93%	14% / 50%	97%	97% / 97%	97% / 97%
Short Segments (≤ 2)	none	61% / 66%	62% / 69%	81%	83% / 85%	76% / 81%
Short Segments (≤ 5)	none	65% / 67%	59% / 64%	67%	73% / 75%	65% / 68%
Raw Crawl Data		94% / 95%	60% / 63%	84%	93% / 94%	79% / 84%
Overtranslation	src	67% / 67%	62% / 68%	52%	66% / 66%	62% / 68%
Undertranslation	trg	69% / 70%	64% / 70%	50%	68% / 68%	63% / 70%

Figure 14: De→En Task: Accuracy of filtering methods with respect to different noise categories

Noise Category	Corrupted Side	Filtering Accuracy				
		Cross Entropy	LASER	Language ID Filtering +none	+ CE	+ LASER
Misaligned Sentences	none	71% / 71%	72% / 72%	50%	62% / 65%	61% / 66%
Misordered Words	src	63% / 64%	53% / 54%	50%	57% / 62%	51% / 53%
	tgt	84% / 84%	50% / 51%	50%	69% / 76%	51% / 51%
Untranslated	src	69% / 70%	4% / 50%	86%	86% / 86%	86% / 86%
	trg	93% / 93%	2% / 50%	86%	86% / 86%	86% / 86%
Raw Crawl Data		77% / 77%	40% / 50%	71%	71% / 77%	70% / 71%
Overtranslation	src	56% / 56%	54% / 55%	51%	53% / 55%	52% / 54%
Undertranslation	trg	63% / 63%	61% / 61%	50%	58% / 60%	56% / 59%

Figure 15: Km→En Task: Accuracy of filtering methods with respect to different noise categories

- Output of pooling strategy is feed forwarded to Softmax layer.

- **STransQuest:**
  - Original text and its translation are fed to two different XLM-R models.
  - Cosine Similarity is computed between the pooling layer's output.

The three pooling strategies of transformer model are CLS, Max, Mean. The objective function used is MSE Loss.

Experiment Results

The results of Domain Adaptation scores are shown in the figure 17.

9 Automatic Post-Editing

Automatic Post-Editing (APE) is a supplementary task within the field of Machine Translation (MT) that focuses on the automatic identification and correction of errors present in MT output (Chatterjee et al., 2020). APE systems have the potential to reduce human effort by correcting systematic and repetitive translation errors (Läubli et al., 2013; Pal et al., 2016). Recent APE approaches utilize transfer learning by adapting pretrained language or translation models to perform APE (Lopes et al., 2019; Wei et al., 2020; Sharma et al., 2021). Also, the recent approaches use multilingual or cross-lingual models to get latent repre-

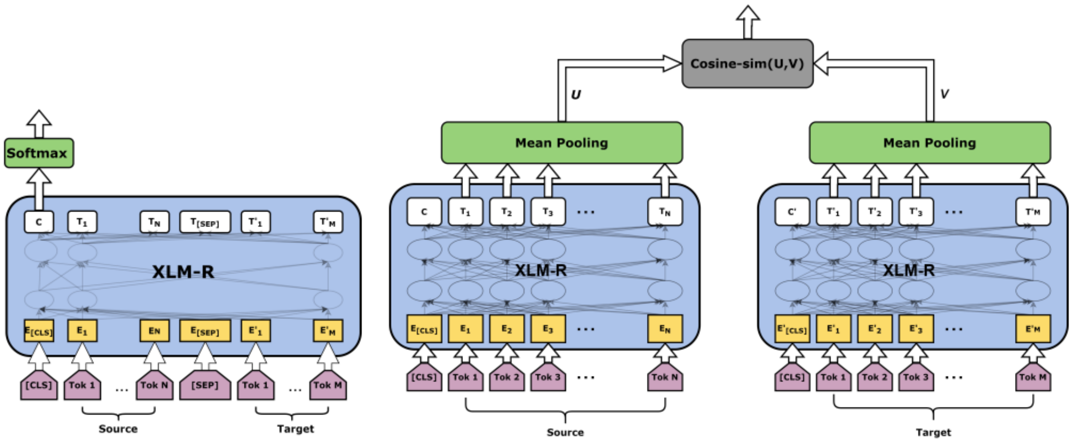

Figure 16: (left) MTransQuest Architecture. (right) STransQuest Architecture.

	Method	Low-resource		Mid-resource			High-resource	
		Si-En	Ne-En	Et-En	Ro-En	Ru-En	En-De	En-Zh
I	MTransQuest	0.6525	**0.7914**	0.7748	**0.8982**	0.7734	**0.4669**	**0.4779**
	STransQuest	0.5957	0.7081	0.6804	0.8501	0.7126	0.3992	0.4067
II	MTransQuest *-En|En-*	**0.6528**	0.7824	**0.7827**	0.8868	**0.7821**	0.4518	0.4334
	STransQuest *-En|En-*	0.5968	0.6992	0.6921	0.8432	0.7152	0.3621	0.3812
III	MTransQuest-m	0.6526	0.7581	0.7574	0.8856	0.7521	0.4420	0.4646
	STransQuest-m	0.5970	0.6980	0.6934	0.8426	0.6945	0.3832	0.3900
IV	OpenKiwi	0.3737	0.3860	0.4770	0.6845	0.5479	0.1455	0.1902
	TransQuest @WMT2020	0.6849	0.8222	0.8240	0.9082	0.8082	0.5539	0.5373
V	mBERT	NS	0.6452	0.6231	0.8351	0.6661	0.3765	0.3982

Figure 17: Correlation between TransQuest predictions and human annotated DA scores

sentations of the source and target sentences (Lee et al., 2020). Oh et al. (2021) have shown that gradually adapting pre-trained models to APE by using the Curriculum Training Strategy (CTS) improves performance. Deoghare and Bhattacharyya (2022) showed that augmenting the APE data with phrase-level APE triplets improves feature diversity, and using a QE system allows for identification and discarding poor-quality APE outputs. We use the APE system to rectify errors in the target side of the noisy pseudo-parallel corpus.

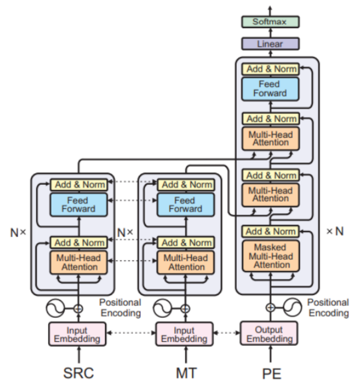

Figure 18: Automatic Post-Editing model Architecture

Model Architecture

A curriculum training strategy for training the APE (Automatic Post-Editing) system was introduced by Deoghare and Bhattacharyya (2022) (Deoghare and Bhattacharyya, 2022). We adopt the same approach to train our APE system. Initially, we employ a pseudo-parallel corpus comprising Samanantar, Anuvaad, ILCI, and the Tatoeba corpus to train an encoder-decoder model specifically for English-to-Marathi translation. Subsequently, we enhance the model by introducing an additional encoder, resulting in a dual-encoder single-decoder model specifically designed for the APE task. This train-

Source	en-as	en-bn	en-gu	en-hi	en-kn	en-ml	en-mr	en-or	en-pa	en-ta	en-te	Total
Existing Sources	108	3,496	611	2,818	472	1,237	758	229	631	1,456	593	12,408
New Sources	34	5,109	2,457	7,308	3,622	4,687	2,869	769	2,349	3,809	4,353	37,366
Total	141	8,605	3,068	10,126	4,094	5,924	3,627	998	2,980	5,265	4,946	49,774
Increase Factor	1.3	2.5	5	3.6	8.7	4.8	4.8	4.4	4.7	3.6	8.3	4

Figure 19: Samanantar Data Statistics

Mykhel	DD national + sports	Punjab govt	Pranabmukherjee	Catchnews	Nptel
Drivespark	Financial Express	Gujarati govt	General_corpus	Kolkata24x7	Wikipedia
Good returns	Zeebiz	Business Standard	NewsOnAir	Asianetnews	Coursera
Indian Express	Sakshi	The Wire	Nouns_dictionary	YouTube science channels	
The times of india	Marketfeed	The Bridge	PIB	Prothomalo	
Nativeplanet	Jagran	The Better India	PIB_archives	Khan_academy	

Figure 20: Samanantar Machine Readable sources

ing process involves multiple stages, incorporating synthetic APE data, and finally fine-tuning the model using real APE data.

10 Comparable Corpora

A comparable corpus is a collection of similar sentences in multiple languages that are not necessarily aligned. For instance, sentences crawled for Wikipedia's mulitlingual pages. Such sentences need not be exact translations of each other or aligned but they refer to the same topic in different languages. In this chapter, we study the work presented in (Ramesh et al., 2021). The work aimed to compile the largest open-source parallel corpora for Indian languages.

A total of 49.7M parallel sentences were collated between English and 11 Indic languages. The web-crawled corpora were of size 37.4M sentence pairs. They also extracted 53.4M sentence pairs between all 55 Indian languages. The data statistics of collated corpora is shown in fig 19

The mining of parallel sentences from the web was achieved by combining various corpora, tools, and methods:

- Web-crawled monolingual corpora
- Extraction from scanned documents was performed by using a document OCR
- Multilingual sentence embedding models for sentence alignment

The quality of the Samanantar Corpus was verified by training a multilingual model on the collected corpus and comparing its BLEU scores against the state-of-the-art models.

10.1 Samanantar Corpus

Samanantar is the largest publicly available corpora collection for Indic languages. It contains datasets for languages like Assamese, Malayalam, Marathi, Oriya, Punjabi, Bengali, Gujarati, Hindi, Kannada, Tamil, Telugu, and English. It has 49.6M sentence pairs between English to Indic languages. The various methods used to collect parallel corpora and build the Samanantar Corpus are mentioned below.

10.2 Collation from existing resources

A total of 12.4M parallel sentences are collected between English and 11 Indic languages from the existing resources. However, some of these resources were very noisy and combined without qualitative filtering.

Mining Sentences from Machine Readable Comparable Corpora

Comparable Corpora are extracted from Indian news articles published in multiple languages. These articles are considered comparable because although there are no exact translations of each other but they are on the same topic. For instance, a news article for COVID'19 published in multiple sentences may not be exact translations of each other, but there can exist some unaligned parallel sentences. Some comparable corpora is also crawled from education domains like Khan

	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	Total
as	-	356	142	162	193	227	162	70	108	214	206	1839
bn		-	1576	2627	2137	2876	1847	592	1126	2432	2350	17920
gu			-	2465	2053	2349	1757	529	1135	2054	2302	16361
hi				-	2148	2747	2086	659	1637	2501	2434	19466
kn					-	2869	1819	533	1123	2498	2796	18168
ml						-	1827	558	1122	2584	2671	19829
mr							-	581	1076	2113	2225	15493
or								-	507	1076	1114	6218
pa									-	1747	1756	11336
ta										-	2599	19816
te											-	20453

Figure 21: Samanantar Machine Readable sources

Academy, NPTEL lectures, Coursera and some science youtube channels. After the collection of comparable corpora, parallel sentences are aligned using LaBSE. For instance, a Hindi and English news article with the same headline published on the same date is taken. Now sentence embeddings are computed for each sentence in the articles for both languages using LaBSE. Then, sentence pairs are aligned using the cosine similarity computed using their respective sentence embeddings. A list of machine-readable sources is shown in fig 20.

Mining sentences from non-machine readable comparable corpora

Apart from web sources, there exist non-machine readable sources like PDF documents. For such documents, OCR tool is used to extract text from the PDF. These documents are available with their language information. Hence, Parallel corpus extraction for such documents becomes easy as we just need to map sentences between different languages.

Mining from monolingual corpora

In this method, parallel sentences are aligned and extracted from the IndicCorp dataset. The idea is to find a matching English sentence for each Indic sentence. Firstly, sentence embeddings are generated for each sentence using LaBSE. Then, FAISS is used for indexing. Now, for each Indic sentence, LaBSE sentence embedding is computed, and then based on the normalized inner product, the index is queried for its nearest neighbor.

Mining Corpora between Indic launguages

In this method, English is used as the pivot language to extract parallel corpora between Indic languages from the mined English-centric corpora. For instance, sentence pairs from English-Hindi and English-Tamil are mapped to each other if the source English sentence is the same. This way, we obtain parallel Hindi-Tamil corpora. The data statistics for the parallel corpora between Indic languages is shown in fig 21.

10.3 Experiment Results

Annotation Task

Human annotators analyzed the quality of the mined corpora by estimating the Sentence Text Similarity of the mined parallel sentences. 9,566 parallel sentences are sampled from the mined corpora of 11 Indic languages. Annotation scores follow the SemEval-2016 Task 1, where STS is defined by six levels i.e. 0-6, 6 being completely semantic equivalent and 0 being entirely semantic dissimilar. The annotation results are shown in fig 22.

IndicTrans

The multilingual NMT model is trained on the entire Samanantar corpus using OpenNMT-py. The results of the trained model are shown in fig 23 and 24. The IndicTrans model outperforms all publicly

Language	Annotation data		Semantic Textual Similarity score				Spearman correlation coefficient		
	# Bitext pairs	# Anno-tations	All accept	Definite accept	Marginal accept	Reject	LAS, STS	LAS, Sentence len	STS, Sentence len
Assamese	689	1,972	3.52	3.86	3.11	2.18	0.25	-0.39	0.19
Bengali	957	3,797	4.59	4.86	4.31	3.53	0.45	-0.43	-0.16
Gujarati	779	2,298	4.08	4.54	3.59	2.67	0.49	-0.31	-0.08
Hindi	1,276	4,616	4.50	4.84	4.14	3.15	0.48	-0.18	-0.12
Kannada	957	2,838	4.20	4.61	3.78	2.81	0.39	-0.38	-0.09
Malayalam	948	2,760	4.00	4.46	3.55	2.45	0.40	-0.33	0.03
Marathi	779	1,984	4.07	4.52	3.54	2.67	0.40	-0.36	-0.04
Odia	500	1,264	4.49	4.63	4.34	4.33	0.15	-0.42	-0.05
Punjabi	688	2,222	4.23	4.67	3.74	2.32	0.43	-0.25	0.06
Tamil	1,044	2,882	4.29	4.62	3.95	2.57	0.35	-0.40	-0.14
Telugu	949	2,516	4.62	4.87	4.34	3.62	0.36	-0.40	-0.09
Overall	9,566	29,149	4.27	4.63	3.89	2.94	0.37	-0.35	-0.04

Figure 22: Samanantar Machine Readable sources

	x-en									en-x								
Model	GOOG	MSFT	CVIT	OPUS	mBART	TF	mT5	IT	Δ	GOOG	MSFT	CVIT	OPUS	mBART	TF	mT5	IT	Δ
								WAT2021										
bn	20.6	21.8	-	11.4	4.7	24.2	24.8	**29.6**	4.8	7.3	11.4	12.2	-	0.5	13.3	13.6	**15.3**	1.7
gu	32.9	34.5	-	-	6.0	33.1	34.6	**40.3**	5.7	16.1	22.4	22.4	-	0.7	21.9	24.8	**25.6**	0.8
hi	36.7	38.0	-	13.3	33.1	38.8	39.2	**43.9**	4.7	32.8	34.3	34.3	11.4	27.7	35.9	36.0	**38.6**	2.6
kn	24.6	23.4	-	-	-	23.5	27.8	**36.4**	8.6	12.9	16.1	-	-	-	12.1	17.3	**19.1**	1.8
ml	27.2	27.4	-	5.7	19.1	26.3	26.8	**34.6**	7.3	10.6	7.6	11.4	1.5	1.6	11.2	7.2	**14.7**	3.3
mr	26.1	27.7	-	0.4	11.7	26.7	27.6	**33.5**	5.9	12.6	15.7	16.5	0.1	1.1	16.3	17.7	**20.1**	2.4
or	23.7	27.4	-	-	-	23.7	-	**34.4**	7.0	10.4	14.6	16.3	-	-	14.8	-	**18.9**	2.6
pa	35.9	35.9	-	8.6	-	36.0	37.1	**43.2**	6.1	22	28.1	-	-	-	29.8	31.	**33.1**	2.1
ta	23.5	24.8	-	-	26.8	28.4	27.8	**33.2**	4.8	9.0	11.8	11.6	-	11.1	12.5	13.2	**13.5**	0.3
te	25.9	25.4	-	-	4.3	26.8	28.5	**36.2**	7.7	7.6	8.5	8.0	-	0.6	12.4	7.5	**14.1**	1.7
								WAT2020										
bn	17.0	17.2	18.1	9.0	6.2	16.3	16.4	**20.0**	1.9	6.6	8.3	8.5	-	0.9	8.7	9.3	**11.4**	2.1
gu	21.0	22.0	23.4	-	3.0	16.6	18.9	**24.1**	0.7	10.8	12.8	12.4	-	0.5	9.7	11.8	**15.3**	2.5
hi	22.6	21.3	23.0	8.6	19.0	21.7	21.5	**23.6**	0.6	16.1	15.6	16.0	6.7	13.4	17.4	17.3	**20.0**	2.6
ml	17.3	16.5	18.9	5.8	13.5	14.4	15.4	**20.4**	1.5	5.6	5.5	5.3	1.1	1.5	5.2	3.6	**7.2**	1.6
mr	18.1	18.6	19.5	0.5	9.2	15.3	16.8	**20.4**	0.9	8.7	10.1	9.6	0.2	1.0	9.8	10.9	**12.7**	1.8
ta	14.6	15.4	17.1	-	16.1	15.3	14.9	**18.3**	1.3	4.5	5.4	4.6	-	5.5	5.0	5.2	**6.2**	0.7
te	15.6	15.1	13.7	-	5.1	12.1	14.2	**18.5**	2.9	5.5	7.0	5.6	-	1.1	5.0	5.4	**7.6**	0.7
								WMT										
hi	31.3	30.1	24.6	13.1	25.7	25.3	26.0	**29.7**	-1.6	24.6	24.2	20.2	7.9	18.3	23.	23.8	**25.5**	0.9
gu	30.4	29.9	24.2	-	5.6	16.8	21.9	**25.1**	-5.4	15.2	17.5	12.6	-	0.5	9.0	12.3	**17.2**	-0.3
ta	27.5	27.4	17.1	-	20.7	16.6	17.5	**24.1**	-3.4	9.6	10.0	4.8	-	6.3	5.8	7.1	**9.9**	-0.1
								UFAL										
ta	25.1	25.5	19.9	-	24.7	26.3	25.6	**30.2**	3.9	7.7	10.1	7.2	-	9.2	11.3	**11.9**	10.9	-1.0
								PMI										
as	-	16.7	-	-	-	7.4	-	**29.9**	13.2	-	10.8	-	-	-	3.5	-	**11.6**	0.8

Figure 23: Samanantar Machine Readable sources

	x-en							en-x						
Model	GOOG	MSFT	CVIT	OPUS	mBART	IT†	IT	GOOG	MSFT	CVIT	OPUS	mBART	IT†	IT
as	-	24.9	-	-	-	17.1	**23.3**	-	13.6	-	-	-	**7.0**	6.9
bn	34.6	31.2	-	17.9	9.4	30.1	**32.2**	28.1	22.9	7.9	-	1.4	18.2	**20.3**
gu	40.2	35.4	-	-	4.8	30.6	**34.3**	25.6	27.7	14.1	-	0.7	19.4	**22.6**
hi	44.2	36.9	-	18.6	32.6	34.3	**37.9**	38.7	31.8	25.7	13.7	22.2	32.2	**34.5**
kn	32.2	30.5	-	-	-	19.5	**28.8**	32.6	22.0	-	-	-	9.9	**18.9**
ml	34.6	34.1	-	9.5	24.0	26.5	**31.7**	27.4	21.1	6.6	4.4	3.0	10.9	**16.3**
mr	36.1	32.7	-	0.6	14.8	27.1	**30.8**	19.8	18.3	8.5	0.1	1.2	12.7	**16.1**
or	31.7	31.0	-	-	-	26.1	**30.1**	24.4	20.9	7.9	-	-	11.0	**13.9**
pa	39.0	35.1	-	9.9	-	30.3	**35.8**	27.0	28.5	-	-	-	21.3	**26.9**
ta	31.9	29.8	-	-	22.3	24.2	**28.6**	28.0	20.0	7.9	-	8.7	10.2	**16.3**
te	38.8	37.3	-	-	15.5	29.0	**33.5**	30.6	30.5	8.2	-	4.5	17.7	**22.0**

Figure 24: Samanantar Machine Readable sources

available models. It also outperforms commercial models on many datasets.

11 Dataset

The parallel and pseudo-parallel corpora consist various datasets. The description for these datasets is mentioned below:

1. **Samanantar[1]:** Samanantar is the largest publicly available corpora collection for Indic languages. It contains datasets for languages like Assamese, Malayalam, Marathi, Oriya, Punjabi, Bengali, Gujarati, Hindi, Kannada, Tamil, Telugu, and English. It has 49.6M sentence pairs between English to Indic languages.

2. **ILCI[2]:** Indian Language Corpora Initiative was envisioned by TDIL to develop national corpora. The ILCI phase-1 contains parallel annotated corpora for 12 major Indian languages including English. It contains sentence pairs from Healthcare and Tourism domain.

3. **PMIndia[3]:** PMI is a good quality publicly available parallel corpora which cover 13 major Indian languages with English.

4. **CVIT-PIB[4]:** PIB consists of Parallel text between English and 9 Indic languages extracted by aligning and mining parallel sentences

[1] https://ai4bharat.iitm.ac.in/samanantar

[2] http://sanskrit.jnu.ac.in/projects/ilci.jsp?proj=ilci

[3] https://www.kaggle.com/datasets/taruntiwarihp/pm-india-mann-ki-baat

[4] https://pib.gov.in

from press releases of the Press Information Bureau of India.

5. **Bible**[5]**:** It is multilingual parallel corpora created from the translations of the Bible. It covers 102 languages including Indian languages.

6. **Tatoeba**[6]**:** The Tatoeba Translation Challenge dataset contains train and test data for 500 languages.

7. **Paramed**[7]**:** This corpus consists of parallel sentences of the biomedical domain for English-Chinese.

8. **GNOME**[8]**, KDE4**[9]**, Ubuntu**[10]**:** They consist sentence pairs between 11 Indic languages and English in their respective localization.

9. **OPUS**[11]: It is an opensource repository for webcrawled text. We use all the parallel data present at OPUS for the Hindi-Bengali language pair

12 Summary

In this survey paper, we first discussed different approaches proposed for the task of Parallel Corpus Filtering. We studied Parallel Corpus Filtering in Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). Then, we discussed different ways to construct comparable corpora. We also discussed various datasets used for the task of Machine Translation.

References

P. Bhattacharyya. 2015. *Machine Translation*. A Chapman Hall book. CRC Press, Taylor Francis Group.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

[5] https://opus.nlpl.eu/bible-uedin.php
[6] https://github.com/Helsinki-NLP/Tatoeba-Challenge
[7] https://github.com/boxiangliu/ParaMed
[8] https://l10n.gnome.org/
[9] https://l10n.kde.org/
[10] https://translations.launchpad.net/
[11] https://opus.nlpl.eu/

Sourabh Deoghare and Pushpak Bhattacharyya. 2022. Iit bombay's wmt22 automatic post-editing shared task submission. In *Proceedings of the Seventh Conference on Machine Translation*, pages 682–688, Abu Dhabi. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Samuel Läubli, Mark Fishel, Gary Massey, Maureen Ehrensberger-Dow, and Martin Volk. 2013. Assessing post-editing efficiency in a realistic translation environment. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*, Nice, France.

Jihyung Lee, WonKee Lee, Jaehun Shin, Baikjin Jung, Young-Kil Kim, and Jong-Hyeok Lee. 2020. POSTECH-ETRI's submission to the WMT2020 APE shared task: Automatic post-editing with cross-lingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 777–782, Online. Association for Computational Linguistics.

António V. Lopes, M. Amin Farajian, Gonçalo M. Correia, Jonay Trénous, and André F. T. Martins. 2019. Unbabel's submission to the WMT2019 APE shared task: BERT-based encoder-decoder for automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 118–123, Florence, Italy. Association for Computational Linguistics.

Shinhyeok Oh, Sion Jang, Hu Xu, Shounan An, and Insoo Oh. 2021. Netmarble AI center's WMT21 automatic post-editing shared task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 307–314, Online. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, and Josef van Genabith. 2016. Multi-engine and multi-alignment based automatic post-editing and its impact on translation productivity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2559–2570, Osaka, Japan. The COLING 2016 Organizing Committee.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Abhishek Sharma, Prabhakar Gupta, and Anil Nelakanti. 2021. Adapting neural machine translation for automatic post-editing. In *Proceedings of the Sixth Conference on Machine Translation*, pages 315–319, Online. Association for Computational Linguistics.

Inguna Skadiņa, Ahmet Aker, Nikos Mastropavlos, Fangzhong Su, Dan Tufis, Mateja Verlic, Andrejs Vasiļjevs, Bogdan Babych, Paul Clough, Robert Gaizauskas, Nikos Glaros, Monica Lestari Paramita, and Mārcis Pinnis. 2012. Collecting and using comparable corpora for statistical machine translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Daimeng Wei, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, and Shiliang Sun. 2020. HW-TSC's participation in the WMT 2020 news translation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 293–299, Online. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.