Automatic Resource Augmentation for Machine Translation in Low Resource Language: `EnIndic Corpus`

ANASUA BANERJEE, National Institute of Technology, Jamshedpur, India
VINAY KUMAR, National Institute of Technology, Jamshedpur, India
ACHYUT SHANKAR, WMG, University of Warwick, United Kingdom
RUTVIJ H. JHAVERI, Pandit Deendayal Energy University, India
DEBAJYOTY BANIK, **(Corresponding author)** Member, ACM, School of Computer Science and Artificial Intelligence, SR University, India

Parallel corpus is the primary ingredient of machine translation. It is required to train the statistical machine translation (SMT) and neural machine translation (NMT) systems. There is a lack of good quality parallel corpus for Hindi to English. Comparable corpora for a given language pair are comparatively easy to find, but this cannot be used directly in SMT or NMT systems. As a result, we generate a parallel corpus from the comparable corpus. For this purpose, the sentences (which are translations of each other) are mined from the comparable corpus to prepare the parallel corpus. The proposed algorithm uses the length of the sentence and word translation model to align sentence pairs that are translations of each other. Then, the sentence pairs that are poor translations of each other (measured by a similarity score based on IBM model 1 translation probability) are filtered out. We apply this algorithm to comparable corpora, which are crawled from speeches of the President and Vice-President of India, and mined parallel corpora out of them. The prepared parallel corpus contains good quality aligned sentences (with 96.338% f-score). Subsequently, incorrect sentence pairs are filtered out manually to make the corpus in qualitative practical use. Finally, we gather various sentences from different sources to prepare the EnIndic corpus, which comprises 1,656,207 English-Hindi sentence pairs (miscellaneous domain). We have deployed this prepared largest English-Hindi parallel corpus at https://github.com/debajyoty/EnIndic.git and the source code at https://github.com/debajyoty/EnIndicSourceCode.git.

CCS Concepts: • **Computing methodologies → Machine translation**.

Additional Key Words and Phrases: Parallel Corpus, Comparable Corpus, Machine Translation, Linguistic Resources and Natural Language Processing

ACM Reference Format:
Anasua Banerjee, Vinay Kumar, Achyut Shankar, Rutvij H. Jhaveri, and Debajyoty Banik. 2022. Automatic Resource Augmentation for Machine Translation in Low Resource Language: `EnIndic Corpus`. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 37, 4, Article 111 (August 2022), 17 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Secure IoMT illness prediction utilizing Multimedia Information Processing techniques has been proposed by Ghazal et al. [19] using the Private Blockchain and Fuzzy Logic based Attack Detection System (PBFL-ADS). To find pertinent knowledge by examining connections between Web of Things data using decomposition strategies, they introduce a novel method called Decomposition

Authors' addresses: Anasua Banerjee, anasua123.banerjee@gmail.com, National Institute of Technology, Jamshedpur, Jamshedpur, India, 831014; Vinay Kumar, vkumar.cse@nitjsr.ac.in, National Institute of Technology, Jamshedpur, Jamshedpur, India, 831014; Achyut Shankar, ashankar2711@gmail.com, WMG, University of Warwick, Coventry, United Kingdom; Rutvij H. Jhaveri, rutvij.jhaveri@sot.pdpu.ac.in, Pandit Deendayal Energy University, India; Debajyoty Banik, debajyoty.banik@gmail.com **(Corresponding author)**, Member, ACM, School of Computer Science and Artificial Intelligence, SR University, Warangal, Telangana, India, 506371.

2375-4699/2022/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

for Ontology Matching (DOM), which focuses on semantic modeling of the Web of Things [9]. In machine translation (MT), many research works are going on, and there is quite a development in this field. It can quickly reduce the language barrier among people of different geographical locations and cultures. Day by day, content on the Internet is increasing. However, its reachability is limited as most of the contents are still in English, and this causes a language inconvenient for people of other languages. Due to its population, we consider the content in the context of the Indian point of view. The central government's official languages are Hindi and English, and the Indian Constitution recognizes twenty-two languages as scheduled. This diversity creates many problems in India, especially for the Government of India. They must perform a manual translation from one language to another to solve the issues. It is a very costly operation. Taking help from the automatic systems to reduce the underlying cost. The authors [16] surveyed, that the presentation of dual-language reading materials has a greater effect on students' outcomes than the presentation of reading materials in only one type of language for students with an intermediate language level. In this paper, [3] the authors gave a succinct overview of the methods currently in use for event identification from shared social multimedia data on the Web, contrasting each method's features in light of some difficulties like collective knowledge, misleading content, unstructured content, etc.

The accuracy of translation and research systems has recently increased dramatically due to hybrid machine translation methods. Additionally, there is always a possibility that some translated outputs produced by one system may be superior to the corresponding translated outputs produced by another system, whereas the situation may be the reverse for the remaining portion of the sentence. To maximize the performance, the authors [8] applied a system combination approach. For unsupervised machine translation techniques to perform better, parallel sentence mining and cleaning are essential tasks. Here, we tried to mine the parallel Hindi-English dataset from a comparable corpus by employing an algorithm that aligns phrase pairs that are translations of one another based on word and sentence length. The sentence pairs that are poor translations of one another are then eliminated (based on a similarity value derived from the IBM model 1 translation probability).

NMT and SMT systems automatically learn the pattern of translations from the parallel corpus to build the model for translation. The bilingual parallel corpus should be of good quality and sufficiently large to train the translation system. This is the primary ingredient of natural language processing (NLP) or artificial intelligence (AI) research.

Finding such a large and good-quality bilingual parallel corpora for a given language pair is not easy. But, comparable corpora are comparatively easy to find out. The parallel corpus is sentence-aligned multilingual data. The translation accuracy also depends on the quality and quantity of the parallel corpus. The comparable corpus is text pair that come from the same domain in different languages but are neither sentence aligned nor of the same size. Comparable corpora can be the same news article or similar Wikipedia article in other languages. Many bilingual resources are comparable but not exactly translations of each other, i.e., news articles, Wikipedia articles, etc. We cannot use the comparable corpora directly for MT system training. So, we need to convert the comparable corpus into the sentence-aligned parallel corpus. Indian languages are not resourceful languages, and there is a lack of resources to work with for machine translation or any other computational linguistic application. Some good quality parallel corpora of various Indian languages have been created under the Indian Language Corpora Initiative (ILCI) [23]. These corpora are good, but the size is minimal for machine translation training. Therefore, there is a need to prepare more good quality parallel corpora of these languages to train machine translation systems successfully. There is abundant text available in Hindi and English, which are comparable. These comparable corpora can be mined, and the sentence-aligned parallel corpora can be prepared using the proposed approach.

There are some existing algorithms to extract parallel sentences from comparable bilingual corpora. We will present a new algorithm to do the same with higher accuracy and relatively faster than existing methodologies. Some existing methods of sentence alignment, such as sentence length-based and word correspondence-based, are incorporated together to achieve a better alignment in our proposed algorithm. Then, the algorithm filters out the poor-quality translation pairs from aligned sentences using a similarity measure based on the sentence pairs' IBM Model 1 translation score.

Generally, an algorithm to create parallel corpora from a comparable one can be broken as a two-step process. Firstly, it forms the sentence pairs, expected to be the translation of each other by aligning the sentences of the bilingual corpus. After that, the poor-quality translations are filtered out from the aligned sentence pairs. The first step of alignment is computationally expensive as we should go through the whole corpus of the target language to find the most suitable translation of one sentence of the source language.

A bilingual parallel corpus with a large size and quality for a particular language pair is tedious to find out. Comparable corpora, however, are relatively simple to find. The parallel corpus contains multilingual data that is sentence-aligned. The standard and size of the parallel corpus also affect how accurate the translation is. Development of a parallel corpus from the comparable corpus manually is a time-consuming task. This existing problem invokes us to develop the EnIndic English-Hindi Parallel corpus.

This paper exclusively shows the preparation of a Hindi-English parallel corpus from the comparable corpus whose size is 1,656,207. In the case of IITB English-Hindi Parallel Corpus [27] is 1,492,827 only. There is no language-specific algorithm. Because of this, it might produce parallel corpora using sources of comparable corpora for other specified language combinations. Furthermore, while creating the corpus we determined the threshold value using empirical analysis, which provides the highest level of accuracy.

1.1 Related Work

Data is an essential element for every machine learning-based task [1, 2, 7, 8]. Most of the early efforts focus on sentence lengths for sentence alignment. One of the earliest approaches to aligning sentences is described in the literature [13]. It is a sentence length-based probabilistic approach. The simple word count is the backbone of the whole process. They considered that every bilingual corpus could be aligned as the sequence of "beads", where the bead was a minimal alignment segment. There were five types of beads assumed: 1-to-1, 1-to-2, 2-to-1, 1-to-0, or 0-to-1. The alignment model was a probabilistic generative model that tried to predict the length of a sentence made up of successive beads. Each bead in the sequence was assumed to be generated according to a fixed probability distribution over bead types. There is another model that generates the lengths of the sentences composing the bead for each type of bead.

Another length-based approach was given by [18]. The key distinction between this approach and the one described in [13] was that character length was taken into account rather than word count when determining sentence length. This is purely a statistic-based approach where a probabilistic score is assigned depending on the character-based scaled difference between the sentence lengths. The concept of dynamic programming (DP) takes an active role here in locating the maximum likelihood of sentence alignment.

These length-based approaches work well in some cases but, in most cases, do not perform very well due to not having the same meaning as the sentences in the sentence pair. After translation, the target language words can be clubbed to form a single word many times. Sometimes two sentences may have the same sentence length ratio, which aligns them but might not be good translations.

However prior strategies employed these kinds of sentence combinations to align the clauses. Researchers proposed a similarity measure to filter out the poor quality translations from the aligned sentence pairs [42]. Sentence pairs were preliminarily aligned using a similar approach depending on the sentence length-based DP-matching discussed above. The similarity measure $SIM(S_i, T_i)$ depends on various words in the source language sentence (S_i), which has the translation in the target language sentence (T_i) decided by using the bilingual dictionary. They introduced another sentence similarity measure $SIM_2(S_i, T_i)$ in 2007 [43], given by:

$$SIM_2(J_i, E_i) = \frac{2 \times \sum_{j \in J_i} \sum_{e \in E_i} \frac{\delta(j,e)}{deg(j)deg(e)}}{|J_i| + |E_i|} \tag{1}$$

where,
$|J_i|$ = source tokens count in i^{th} alignment
$|E_i|$ = target tokens count in i^{th} alignment
$\delta(j, e)$ = 1 if j and e can be a translation pair 0 otherwise.
$deg(j) = \sum_{e \in E_i} \delta(j, e)$
$deg(e) = \sum_{j \in J_i} \delta(j, e)$
Using this measure, they also calculated the document similarity score. The authors multiplied the sentence similarity score, document similarity score, and the ratio of the number of sentences in both the corpus to get each sentence pair's final score of alignment. Sentences were filtered out from the final parallel corpus based on this score.

Words must be rooted before comparison because there are many morphological forms of the word in different languages. Also, the translation of some groups of words may become a phrase in another language, and the technique mentioned above cannot identify this.

[40] used a first-order linear chain Conditional Random Field (CRF) for sentence alignment within the aligned documents. They adopted the discriminative CRF-based word alignment model [11]. Researchers also proposed a general methodology to mine parallel corpora from multilingual patent documents [32]. They discussed that the length-based approach is not sufficient to consider content similarity. It depends neither on the bilingual dictionary-based method, as it might not deal with new terms of patents, nor on the translation model-based approach as it needs enough data to do the translation. They combined all three to get a good quality parallel corpora. They first aligned the sentences using the same similarity measure $SIM_2(J_i, E_i)$, shown in Equation 1. After alignment, they removed much longer sentences (the selected threshold is a hundred English words) and performed length ratio filtering by discarding sentence pairs exceeding 0.8 to 1.8. The poor quality translations were filtered out from the parallel corpora using translation similarity score P_t. P_t of sentence pairs is given by combining the translation probability (TP) of sentence pairs in both directions. This way, they combined all three different techniques and formed parallel corpora from comparable corpora of patents. The primary challenge of every method is to correct alignment (word, phrase, or sentence) in a reasonable time. We try to address these challenges for better parallel corpus preparation.

Another famous corpus is by Europarl [25]. The detailed methodology of the data preparation technique employed here is exciting. According to Europarl, preparing parallel corpora from comparable corpora involves five steps. At first, they collected the raw data. They showed the collection of raw data from the website of European Parliament proceedings. The data were collected using web crawling. Each file contained the expressions of a single speaker. Crawling this web resource with a web spider was done by starting at an index page and following certain links, dependent on the inclusion and exclusion rules. This work manually identified the sources for parallel corpora and stated the possibility of mining the web for such data [36]. The next step was document alignment. They identified texts belonging to each topic by processing the HTML data. They used pattern matching to identify the speakers. Then, they grouped similar documents. After that, sentence splitting and tokenization were performed. Truecasing of all words helped to eliminate different spellings of words. The sentences were split using one of these algorithms, SATZ [35], Decision Trees [38], or Maximum Entropy [37]. The next step of sentence alignment is the most important. It is usually a complex problem. It was simplified because texts were already available in the paragraph-aligned format. The data were discarded if the paragraph count of a speaker's utterances differed in different languages. The alignment of sentences was done using the Gale, and Church algorithm [18]. This algorithm tried to identify sentences of identical lengths in the sequences, dependent on the word count in the sentence. The sentence-aligned data was stored in a separate file based on the day and language so that sentences with the same sentence number in any two files were mappings of each other. Also, the markup from the document-aligned file was stripped out. Although this method of preparing parallel corpora from comparable corpora is suitable, it is not precise enough. Some problems were not dealt with significantly in the above ways. These methods do not consider the true casing to eliminate different spellings of words. So, it cannot distinguish different spellings (e.g., black can be a color or the name Mr. Black). Semantic alignment of the sentences was another issue. The algorithm did not consider word correspondences between sentences; it just looked at the lengths of the adjacent sentences. We discovered another paper [10] to address the problems mentioned above. This paper used multilingual patents to mine the parallel corpora on a large scale. The extraction of the comparable corpus was almost the same for this paper. The difference was there in the method of sentence alignment. The work also discussed three ways to align sentences along with their drawbacks and rectification. According to this paper, the particular alignment methods were ineffective because the comparable patents were not strictly parallel. The lexicon-based process cannot handle new terminology of patents. Due to not considering the content similarity, the length-based method cannot perform well. Large high-quality training data is the constraint of the translation model-based method. These three approaches are combined for qualitative sentence alignment from comparable patents. The lexicon is used for preliminary sentence mapping between relative patents. The computed similarity score P_d between sentences is the main key element of this step. The mathematical model is shown in Equation 2.

$$P_d\left(S_c, S_e\right) = \frac{\sum_{W_c \in S_c} \sum_{W_e \in S_e} \frac{\gamma(W_c, W_e)}{deg(W_c) deg(W_e)}}{\frac{l_e + l_c}{2}} \tag{2}$$

Where W_e and W_c are word types in English sentence (S_e) and Chinese sentence (S_c) respectively. If W_e and W_c are translation pair in the dictionary then $\gamma\left(W_c, W_e\right) = 1$ shown in Equations 3 and 4. The lengths of S_e and S_c are l_e and l_c shown in Equation 5.

$$deg(W_c) = \sum_{W_e \in S_e} \gamma(W_c, W_e) \tag{3}$$

$$deg(W_e) = \sum_{W_e \in S_c} \gamma(W_c, W_e) \tag{4}$$

Then they used ratio filtering and length filtering to remove some sentence pairs. They further filtered out some of the remaining parallel sentences using the IBM Model-1 learning and computed the translation similarity score (P_t) between sentence pairs using the TP in both directions.

$$p_t\left(S_c, S_e\right) = \frac{log(P(S_e|S_c)) + log(P(S_c|S_e))}{l_c + l_e} \tag{5}$$

The sentence pairs are again filtered out, and P_t is lower than the threshold. Although this method has advantages in terms of sentence alignment, this is not enough for some language pairs such as Hindi-English. HindEnCorp [12] dealt with some of them. They used Hunalign [44] for the sentence alignment step and suggested trying Bleulign and Gargantua. After that, they performed the cleaning and normalized the texts. They removed various non-printable characters. They removed occasional sequences of nuktas which probably serve as a graphical delimiter. For normalization, they introduced periods instead of danda, true casing, etc. After cleaning and normalization, they performed the morphological analysis using Hindi Morphological Analyzer [39]. Even after these steps, certain drawbacks persisted, which caused hindrances in getting good-quality parallel corpora.

The IIT Bombay English-Hindi parallel corpus [27] is the largest Hindi-English similar dataset. Here, authors manually gathered lots of parallel sentences to prepare parallel corpus as there were done minimal works on Hindi-English parallel corpus (i.e., HindiEnCorp [12], [27]).

Hierarchical document encoder for parallel corpus mining is proposed in [20]. Unsupervised methods for parallel corpus mining are described in [29], [28]. Researchers also filtered parallel corpus using the pre-trained model in [45]. Researchers extracted parallel sentences to improve cross-language information retrieval from Wikipedia in [15]. Authors in [6] developed a parallel corpus for English and Akuapem Twi machine translation. After conducting the clustering job, performance is assessed in the measure of accuracy, which is studied by comparing the predicted cluster of an instance with the real one. They applied an unsupervised deep learning-based network on Urdu short text [5]. The authors [4] proposed a method that includes a subset of unstructured data in the training set and employs a similarity-based approach. In the subsequent cycle of the active learning mechanism, the approach then updates the model training using the new training points. A Betalogger model [22] raises awareness of the problem of smartphone hardware sensors exposing the privacy of users. Using language modeling and a dense multi-layer neural network (DMNN), BetaLogger effectively deduces the written text (long or short) on a smartphone keyboard.

Around 7,000 languages are now spoken worldwide, and almost all language pairs lack sufficient parallel corpus to train machine translation models. The difficulty of creating effective translation models when there is a lack of translated training data has drawn more and more attention in recent years. When there is a lack of training data the performance of NMT degrades. In addition to the technical difficulties of learning under sparse supervision, the absence of freely and openly accessible benchmark datasets makes it challenging to assess approaches trained on low-resource/mid-resource language pairs.

In this paper, [30] different approaches are explored like the Contrastive, Typological, and Translation Mining parallel corpus approaches to address the problem of the translations' target language representativeness. The wide range of representativeness techniques they investigated demonstrates that parallel corpus research today is highly aware of the necessity to control for target language representativeness. Lalita et. al.'s [31] main goal is to create a sizable English-Thai collection for translation purposes. They constructed an English–Thai machine translation dataset with over 1 million segment pairs. They prepared their dataset from various sources: news, Wikipedia articles, SMS messages, task-based dialogs, web-crawled data, government documents, and text artificially generated by a pre-trained language model. Authors in [17] observed that the distribution of the perfect and (perfective) past forms across dialects can be better understood by combining parallel corpus and experimental methods. For machine translation tasks, they presented a parallel corpus [41] for two indigenous Mexican languages—Mazatec (maq) and Mixtec (xtn)—in this study. They also used three alternative ways to assess the usefulness of the corpus.

1.2 Key Contributions

The following is a summary of our key contributions.

(1) In this work, an interesting procedure is introduced for automatically preparing the Hindi-English parallel corpus from the comparable corpus. We get a high-quality Hindi-English parallel corpus using the proposed approach in terms of a better sentence alignment strategy.
(2) These corpora can be used in any application requiring English-Hindi parallel corpora, especially for machine translation. These parallel corpora belong to the news domain. The size of these corpora would also be expanded as the number of speeches present on time goes on.
(3) The algorithm is not language-specific. For this reason, it may create parallel corpora from sources of comparable corpora for other given language pairs.
(4) We identified the threshold value through empirical analysis, which gives the maximum accuracy while creating the corpus.
(5) Finally, we have calculated the Precision, Recall, and F-score of the generated corpus.

2 ALIGNING SENTENCE PAIRS & PRUNING OUT

In this work, the alignment of sentence pairs and pruning out is carried out using four-step processes, the first being a sentence length-based alignment, the second, being filtering out the poor-quality translation pairs, the third being a learning word translation model and the fourth being a word correspondence based alignment. All three methods are discussed in detail in the following subsections. The alignment algorithm is similar to the method which is discussed in [33].

2.1 Sentence Length Based Alignment

In this step, the sentences are aligned using the length of sentences (number of words). To perform this alignment, some modifications were made to Brown's algorithm for improving accuracy and speed. [13]'s algorithm is discussed in brief in Section 1.1. Details can be found in [13] Following major changes are done in the existing algorithm.

2.1.1 Change in Assumed Distribution. [13] assumed 1-to-0 and 0-to-1 bead type. Each sentence length is assumed to be distributed according to the model based on the observed distribution of sentence length in the corresponding language's corpus and in other bead types (1-to-1, 2-to-1, and 1-to-2). The source language sentence length is assumed to be distributed in the same way as in the above case. The target language sentences' total length in the bead is assumed to be distributed according to the model based on the source language sentences' total length in the bead. They also assumed that the logarithm of the length ratio of source language sentences to the length of target language sentences is distributed normally with mean μ and σ. Using these above means they calculated $P(l_t|l_s)$, where l_t is the length of the target language's sentence and where l_s is the length of the source language's sentence.

In this work, the sentence length of the target sentence l_t is assumed to follow the Poisson distribution having to mean l_s times ratio r of mean length of sentences of the target language to source language's mean length.

$$P(l_t|l_s) = \frac{e^{l_s r}(l_s r)^{l_t}}{l_t!} \tag{6}$$

The fundamental idea behind this is that each word in a sentence in the source language is translated into the target language for a total of n words, where n has a Poisson distribution and its meaning is determined by the ratio of the mean sentence lengths in the two languages. We do not need to estimate any parameter such as μ and σ in the original Brown's algorithm. It makes this step much faster and experimentally it was also found that the Poisson distribution better fits the data over the Gaussian distribution model.

2.1.2 Optimizing Search Space and Searching Time. Dynamic programming (DP) is a standard approach for solving alignment problems [18]. The cost function plays an active role in the exhaustive search using DP to align. This makes it have a bad time complexity and practically infeasible for a large corpus. Therefore we need to prune the search. By reducing the search space to a fixed-width band across the primary diagonal of the DP-matching matrix, the search is pruned. The heuristic is to choose the location where it appears to be most aligned within the band and is closest to one of the band boundaries. In the unlikely event that the best-estimated alignment does not even approach a positive minimum distance from the band boundaries, it is acknowledged that the best alignment within the band is the best conceivable alignment at that time, and the search is completed. Otherwise, a wider band is produced and the process is repeated.

2.1.3 Enhancement Boost. Authors in [13] estimated the marginal distributions of sentence durations between two languages based on the raw relative frequencies in the corpus. They estimated the probabilities of the lengths of the shorter sentences and smoothed the estimates for the lengths of the longer phrases by fitting them to the tail of a Poisson distribution. Despite this, the likelihood of each recorded length was calculated from the raw relative frequencies. This only affected the estimates for long sentences, which are uncommon in the corpus, thus it shouldn't have a significant impact on the model's performance.
Thus, the preliminary alignment based on sentence length is completed. Now these aligned sentence pairs will be used to learn the word translation model in the next step of alignment.

2.2 Filtering The Poor Quality Translation

After aligning the corpus, we need to filter out the poor-quality translation pairs from this corpus. To filter out such pairs, we need to predict the threshold value to have the quality of prediction. The threshold value (δ) is calculated using fine tuning with the similarity measure [32] which is discussed in Section 1.1. The sentence pair having a lower similarity score than δ will be filtered out as poor-quality translations. This measure referred to as the translation similarity score P_t is given as:

$$P_t(S_s, S_t) = \frac{log(P(S_s|S_t)) + log(P(S_t|S_s))}{l_s + l_t} \tag{7}$$

where $P(S_t|S_s)$ denotes the probability that IBM model 1-based translator will produce S_t in the target language with S_s in source language as input and vice versa for $P(S_s|S_t)$. In Section 2.3, we already covered the IBM Model 1 translation system. Mathematically, $P(S_t|S_s)$ is given by Equation number refibm model1. The translation probabilities of S_s to S_t and S_t to S_s are combined in equation (7). Thus, this model can be used as a good metric to judge the translation quality of the given sentence pair. As $P(S_t|S_s)$ and $P(S_S|S_t)$ both would lie between 0 and 1 (both inclusive), their respective logarithms would be a negative real number except when the probability is 0. We omit such sentence pairs whose $P(S_t|S_s)$ or $P(S_S|S_t)$ is 0. This is following our goal to remove all poor-quality translations, as this case arises only when we have less evidence that a given sentence pair is a translation of each other. Alignments having translation similarity score which is less than the predefined threshold, are filtered out from the final corpus. Figure 1 shows the graph of the accuracy of alignment vs. similarity threshold. The basis for the choice of this threshold is discussed in the next section.

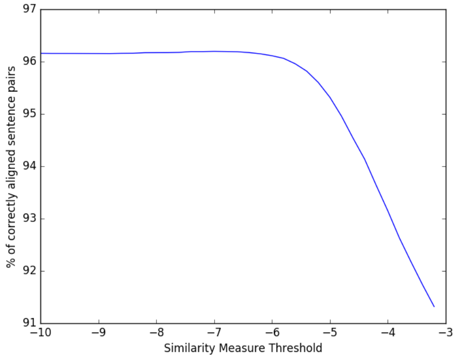

Fig. 1. Graph of alignment accuracy vs. similarity threshold

2.3 Learning Word Translation Model

1-to-1 alignments with a 99% probability of correct alignment are used in this step to train the word translation model using the IBM model 1 translation model with some modifications.

IBM model 1 is a word-based statistical machine translation system, discussed in [14]. It provides the likelihood that a particular sentence s will be translated into t from its source language. In mathematics, it is represented by:

$$p(S_t|S_s) = \frac{\epsilon}{(l_s+1)^{l_t}} \prod_{j=1}^{l_t} \sum_{i=0}^{l_s} t(t_j|s_i) \tag{8}$$

In the following, we discuss some modifications to achieve efficient training of the word translation model.

- The first modification states that due to rarely appearing the rare words, their translation probabilities can be omitted with a negligible amount of loss. As a result, to detect rare words, words are grouped according to how frequently they occur. Finally, before learning the translation probabilities, words with fewer occurrences are mapped into a unique token.
- After the initial modification, the second modification is carried out by accumulating fractional counts for expectation maximization after the first iteration. In a given sentence, a fractional count for a word-translation pair having a lesser value than the predefined threshold ϱ is not part of the translation pair. These ignored fractional counts are assigned to the pair which is formed with the null token. This is feasible because of the original IBM model 1 for all possible pairs consisting of one word from each language. The translation probability of two words, in at least one aligned sentence pair, is calculated and pairs occurring rarely across the whole corpus do not improve quality that much. This modification is done to reduce useless calculations.

The training of this modified version of IBM model 1 is done using the Expectation Maximization (EM) algorithm for four iterations which minimizes the entropy of held data. The word translation model which is formed using this training is used in the next step based on Word-Correspondence.

2.4 Word-Correspondence-Based Alignment

In this final step, the framework of step 1, i.e. length-based approach, is combined with IBM model 1 to perform the alignment. The distribution and generation of bead types and sentence lengths are assumed to be the same as in step 1, except for the probability estimation. It is calculated by the sentence length feature multiplied by an estimated likelihood determined by word correspondence utilizing the word translation score of the IBM model 1. Each word in the corresponding language is believed to be individually created in the 1-to-0 or 0-to-1 bead of each sentence based on the observed relative unigram frequency f_u of the word in the corpus. For all the remaining bead types (1-to-1, 2-to-1, and 1-to-2), the same model of generating words as in the 1-to-0 case is applied to the words of the source language sentence(s). Based on the words in the source language and the instance of Model 1 calculated in the previous phase, it is assumed that the target language sentence(s) in the bead will be formed. Since the alignment model already takes into consideration the element related to the assumption of uniform distribution of the target sentence lengths, it is not necessary to include it when using Model 1. Considering that s and t are, respectively, source and target phrases of lengths l and m. According to the initial model, the probability is $P_{1-1}(l, m)$ to a sentence of length l aligning 1-to-1 with a sentence of length m; then, our combined model estimates the probability of a 1-to-1 bead consisting of s and t as Equation 9.

$$P(s,t) = \frac{P_{1-1}(l,m)}{(l+1)^m}(\prod_{j=1}^{m}\sum_{i=0}^{l} tr(t_j|s_i))(\prod_{i=1}^{l} f_u(s_i)) \tag{9}$$

If this method is used to analyze the entire corpus, it will take a lot of time because it is more expensive than the introduction sentence length-based model. Thus, it is necessary to restrict the search area. Only alignments with a low probability that are produced in the first stage are subject to the approach. As a result, this technique is only used on about 10% of the entire corpus. It greatly aids in accelerating the procedure.

3 EXPERIMENTAL SETUP

3.1 Data preparation

First, we must decide which comparable corpus we should take to generate the parallel corpus. We encountered several sources which act as comparable corpora such as Wikipedia articles and news articles in Hindi and English. During this search, we found that the speeches of the Vice President and President of India were available in both Hindi and English. After going through the lectures, we observed that the translation was pretty accurate, but it is neither sentence-aligned nor all sentences present in both versions. Therefore, these sources can act as a comparable corpus in our research, and good-quality parallel corpora can be mined from them. To gather comparable corpora from speeches of the president and Vice-President of India, we need to collect all the information from their official websites[1] [2]. Speeches delivered by Shri M. Hamid Ansari as Vice-President of India between August 13, 2007, and August 10, 2017, and speeches given by Shri Pranab Mukherjee between August 25, 2012, and July 12, 2017, as President of India, were collected. We crawled over all of their speeches and scrapped all speeches that were available both in Hindi and English. Using Selenium, BeautifulSoup[3] (a python library), this task was done.

Bilingual speeches were kept in a manner that was document-aligned. The Vice President's address contained 24475 English sentences and 22809 Hindi sentences, and the President's speech had 29714 English sentences and 28684 Hindi sentences. Therefore, the total size of the comparable corpus is 54189 sentences in the English corpus and 51493 sentences in the Hindi corpus. We call this dataset "dataset-1" in further Sections.

Besides this, we also need to have a comparable corpus for deciding the accuracy of the algorithm

[1] http://vicepresidentofindia.nic.in/

[2] http://presidentofindia.nic.in/

[3] https://pypi.python.org/pypi/beautifulsoup4

Table 1. Statistical description for comparable corpus and parallel corpus

	Comparable Corpus		Parallel Corpus	
#Sentences	69739	65175	30825	30825
#Tokens	1179519	1340903	563184	648738
#Charecters	7085259	1745077	3576559	9019264

easily. To automatically evaluate the results, we need to have a comparable corpus and a prepared parallel corpus. However, such ready-made datasets are not available publicly. So, we took the parallel corpus and created a pseudo-comparable corpus out of it for evaluation. Such kind of experiments were done using the HindEnCorp corpus (a Hindi-English parallel corpus)[12] (it contains 49999 sentence pairs). We added 1000 sentences and removed 2500 sentences randomly into the HindEnCorp corpus for preparing the pseudo-comparable corpus. The total number of added sentences and removed sentences in the HindEnCorp corpus are however not equal. After this, we get a pseudo-comparable corpus having 49999 sentences in the English corpus and 48499 sentences in the Hindi corpus. We label this dataset as dataset-2 and use the same name in further sections.

3.2 Threshold Selection

Before we start our experiment on the datasets, we need to decide the threshold for filtering out the poor-quality translations. We discuss the similarity measure in Section 2.2 for the filtering process i.e. translation similarity score P_t. The choice of threshold can be made by identifying the threshold value which gives the maximum accuracy while creating a parallel corpus. We need to use dataset 2 to create the parallel corpus as we can check its accuracy automatically. So, we performed the experiments on dataset-2 by varying the threshold of similarity measure between -10.0 to -3.2, with a step size of 0.2. We plot % of correctly aligned sentence pairs for different threshold values in Figure 1. After empirical analysis, it is found that assigning -7.0 to the threshold parameter provides the best accuracy of alignment in the final parallel corpus. So -7.0 is chosen as the threshold to filter out poor-quality translations.

3.3 Used Resources and Tools

To scrap and collect all speeches from the respective websites of the President and Vice-President of India, we used the beautiful soup library of python3. For the bilingual dictionary, we used English-Hindi bilingual mapping [4] which contains many Hindi translations for each English word. It contains 157975 English-Hindi word translation pairs.

4 PREPARED CORPUS STATISTICS AND RESULTS

The prepared corpora consist of a news domain based on 30,825 English-Hindi parallel sentences. The detailed statistics are shown in Table 1. This corpus is split into three segments: training, development, and test set for the building of a standard machine translation system. There are 28,823, 1,001, and 1,001 sentences in the training, development, and test corpora, respectively. The descriptions of these sets are present in Table 2.

By dividing correctly matched sentence pairs by the total number of sentence pairs in the final corpus, precision is calculated. On the other hand, recall is calculated as the ratio of the total number of perfectly aligned sentence pairs in the ideal scenario to the total number of correctly aligned sentence pairs in the final corpus. The high recall is desired when we want a larger corpus with

[4] http://www.cfilt.iitb.ac.in/ sudha/bilingual_mapping.tar.gz

Table 2. Statistics of prepared corpus for the machine translation perspective

	Language	Training	development	Test
#Sentences		28823	1001	1001
#Tokens	English	520479	21104	21601
	Hindi	600166	24104	24468
#Charecters	English	3307582	133248	135729
	Hindi	8356069	329770	333425

Table 3. Manual evaluation for dataset-1 (randomly picked up 1k sentence pair from prepared corpus) after aligning sentences using the hybrid approach applying with P_t for filtering

Methods	Comparable Corpora		Total Number of Sentences Pairs in Parallel Corpora	Total Number of Correctly Aligned Sentence Pairs	Precision (%)	Recall (%)	F-Score (%)
	Number of English Sentences	Number of Hindi Sentences					
Method-filtering	1767	1670	1000	954	98.50	94.27	96.338
Method-hybrid	1767	1670	1000	980	98.0	92.27	95.049
Model-E [10]	1767	1670	1000	972	97.15	91.5	94.24

Table 4. Results for dataset-2: Method-hybrid is sentence aligning hybrid approach and Method-filtering is the hybrid approach with P_t for filtering

Methods	Comparable Corpora		Total Number of Sentences Pairs in Parallel Corpora	Total Number of Correctly Aligned Sentence Pairs	Precision (%)	Recall (%)	F-Score (%)
	Number of English Sentences	Number of Hindi Sentences					
Method-hybrid	49999	48499	46172	44405	96.173	92.778	94.445
Method-filtering	49999	48499	45674	43936	96.195	91.798	93.945
Model-E [10]	49999	48499	45145	42163	94.342	86.834	90.432

more phrase pairings, and high precision is preferred when we want a more precise corpus.

The methodology mentioned above is applied to both dataset-1 and dataset-2 to mine parallel corpora. A generated parallel corpus from dataset 1 contains 31207 sentences in both English and Hindi corpora. The automatically generated parallel corpus (from dataset 1) is newly created. There is no gold set for this corpus. There is also no way to automatically check this corpus's precision and recall. For this reason, checking is done manually, which is a challenging task, especially for calculating recall. So, despite checking the whole parallel and comparable corpus for calculating recall manually, 1000 random sentences are taken out from the parallel corpus, and an evaluation is done. However, to calculate recall, it is necessary to obtain the equivalent similar corpora for these 1000 sentences. The evaluation is reported in Table 3. Here, the method-hybrid and the method-filtering are sentence-aligning hybrid approaches and the hybrid approach with P_t for pruned-out low-scored sentence pairs, respectively.

The evaluated precision value of the entire parallel corpus is 98.73%. As dataset-2 is a pseudo-comparable corpus, we have an existing parallel corpus (gold set) derived from it. Therefore, the parallel corpus mined from it can be automatically evaluated using that gold set parallel corpus. The evaluated results for dataset-2 are shown in Table 4. According to Table 4, the obtained results using the hybrid approach (Method-hybrid) show good precision and recall compared to other techniques. By removing the low-quality translations based on the similarity measure P_t, precision is anticipated to increase. The expected behavior is captured in Table 4. The precision for the hybrid model by filtering with P_t (Method-filtering) is much higher than the hybrid model

Table 5. For Baseline Systems, Results

		Sources			
Language	Test Data	HindEnCorp	VP Speech	IITB	EnIndic
Hi-En	Testset-I	4.94	10.86	32.65	33.91
Hi-En	Testset-II	4.93	2.93	6.28	13.84
Hi-En	Testset-III	8.17	1.35	50.77	63.31
Hi-En	HindEnTest	5.23	3.53	3.48	7.79
Hi-En	IITBTest	4.33	3.73	6.47	11.78
En-Hi	Testset-I	4.07	4.07	34.57	34.22
En-Hi	Testset-II	5.55	5.55	10.58	16.6
En-Hi	Testset-III	7.36	7.36	46.09	45.9
En-Hi	HindEnTest	5.68	3.05	3.53	8.82
En-Hi	IITBTest	4.57	3.81	5.79	13.73

(Method-hybrid). Precision is increased in Method-filtering but not significantly for dataset-2. This type of behavior is not found for dataset-1 (Table 3). The hybrid method, along with filtering by P_t (Method-filtering), provides better alignment than the simple hybrid method (Method-hybrid). Each has better alignment accuracy than the state-of-the-art method [10]. Based on the translation score, the word correspondence model based on IBM model 1 is already applied to preliminary filter out poor quality sentence pairs. In this approach, if the precision of the final corpus is increased, then recall of the final corpus is decreased because some sentence pairs from the final corpus are wrongly filtered out, which should be part of this corpus. Additional filtering based on the translation similarity score should be done when a much higher accuracy of the corpus is required (i.e., accuracy matters more than the size of the corpus). If a large corpus with some lower accuracy is required, the filtering step is optional.

For dataset-1, we reached 98.76% accuracy for the final corpus. The best-aligned corpus leads to a better machine translation system. Instead of using this prepared corpus in training machine translation systems or any other practical usage, the incorrectly aligned sentence pairs need to be removed from the final corpus. Poor quality aligned pairs are removed manually to make 100% accurate final parallel corpus. This way, the usability of the final parallel corpus is increased.

With the provided corpus and the phrase-based statistical machine translation's [26] default settings, we get an accuracy of 13.61 BLEU points for English to Hindi and 14.72 BLEU points for Hindi to English. We start by tokenizing, true-casing, and deleting large sentences and sentences with a length mismatch above a predetermined ratio from the prepared data. Sets for training and development had already been tokenized. For tokenizing English sentences, we use tokenizer.perl[5] script, and for Hindi sentences, we use indic_NLP_Library[6]. The model is trained with trigram language model with modified Kneser-Ney smoothing [24] using KenLM [21].

We have calculated the blue score values for four different datasets, namely HindEnCorp, VP Speech, IITB, and EnIndic, in this Table 5. This table makes it obvious that the Blue score values from our prepared corpus are far superior to those from other corpora. The primary goal of this work is to create a sizable, high-quality English-Hindi parallel corpus. The prepared good-quality English-Hindi parallel corpus is augmented with other existing parallel corpora to prepare the largest English-Hindi parallel corpus. The augmented resources of the parallel corpus are

[5] https://github.com/moses-smt/mosesdecoder/blob/RELEASE-3.0/scripts/tokenizer/tokenizer.perl

[6] https://bitbucket.org/anoopk/indic_nlp_library

Table 6. Various resources to prepare the parallel corpus

Primary Source	# Sentences	Crawling tool/location
Speeches of President and Vice President of India	30,825	Proposed technique
Intercorp	7500	HindEnCorp [12]
Emille	8900	
TED	39800	
Launchpad	66700	
DanielPipes	6600	
Indic	37700	
Tides	50000	
Other smaller sources	56500	
Gnan-Nidhi Corpus	227,123	IITB English-Hindi Parallel Corpus [27]
Wiki Headlines	32,863	
India Government Corpora	123,360	
Judicial domain corpus-I	5,007	
Judicial domain corpus-II	3,727	
Indic Multi parallel corpus	10,349	
TED Talks	42,583	
Mahashabdkosh: Administrative Domain Definitions	46,523	
Mahashabdkosh: Administrative Domain Examples	46,825	
Mahashabdkosh: Administrative Domain Dictionary	66,474	
Hindi-EnglishLinked Wordnets	175,175	
OpenSubs2013	4,222	
Tatoeba	4,698	
Tanzil	187,080	
KDE4	97,227	
GNOME	145,706	
TDIL-ILCI Corpus	49999	ILCI [23]
Indic Languages Multilingual Parallel Corpus	84,553	WAT 2018 [34]

described in Table 6. There have various issues in the existing parallel corpora. The most common problem with the IIT Bombay English-Hindi parallel corpus is the presence of the unwanted word "Completed" instead of any sentence and incomplete sentence. We have tried to consider the issues before gathering corpora together. We need to add/remove some sentences from those to prepare a qualitative parallel corpus. The EnIndic corpus (training set) description is shown in Table 7. Our development set consists of Indic Languages Multilingual Parallel Corpus and IITB English-Hindi Parallel Corpus development sets. We have produced Test Set II. Both Test Set I and Test Set III were taken from VP-Speech and IIIT Bombay English-Hindi, respectively. The sentence count in our development set is 1,020. The token count for the development set is 17,683 and 18,950, and 18,950. We keep three test sets for evaluation: Test Set I (1001 sentences, described earlier), Test Set II (1,000 sentences), and Test Set III (1,000 sentences).

Table 7. Detailed statistics of the EnIndic English-Hindi Parallel Corpus (training set).

Sources	#Sentence	#Token	
		English	Hindi
President and vice-president of India	29,824	541,583	624,270
IITB English-Hindi Parallel Corpus including HindEnCorp	1491827	20,663,652	23,117,342
TDIL-ILCI corpus	49,999	855,412	898,748
Indic Languages Multilingual Parallel Corpus	84,557	516,041	598,022
Total	1,656,207	22,576,688	25,238,382

5 CONCLUSIONS AND FUTURE SCOPE

There are not many high-quality parallel corpora available for Hindi to English. Although it is relatively simple to locate comparable corpora for a particular language pair, these cannot be used directly in SMT or NMT systems. For this reason, we tried to mine the parallel corpus. The concept of removing the sentence pairs that were incorrectly aligned from the final parallel corpus is explored in this paper as a means of improving the quality of mining the parallel corpus from a comparable corpus taken from the speeches of the President and Vice President of India. The suggested method can be used to add more parallel sentence pairs from this website to the final parallel corpus. It is also an intriguing truth that the prepared dataset's quality is checked manually. The linguistic problem that was discovered in the data is then carefully fixed.

Finally, to create the largest English-Hindi parallel corpus (1,656,207), we collected parallel sentences from a variety of sources. In a different domain, the final corpus would be available publicly. The structured EnIndic English-Hindi parallel corpus is anticipated to assist the NLP community in developing data-driven machine translation and numerous other uses.

Our proposed methodology can be further improved by improving the threshold selection approach. The selected threshold for pruning out bad alignment is based on empirical analysis. There is no guarantee for always the best fit with the empirical analysis. Instead of an empirical study, we will try to introduce an algorithm to find the best threshold in the recent future. Furthermore, researchers may incorporate our algorithms on top of different language pairs to make other languages resourceful. Another limitation is algorithm should not be applied to a small comparable corpus as there would not be a sufficient amount of sentence pairs to learn the word translation model using modified IBM model 1 as discussed in section 2.3. This issue needs to be resolved so that comparable corpora with fewer sentences can contribute to the final parallel corpora.

REFERENCES

[1] Ahmed Abbasi, Abdul Rehman Javed, Farkhund Iqbal, Zunera Jalil, Thippa Reddy Gadekallu, and Natalia Kryvinska. 2022. Authorship identification using ensemble learning. *Scientific reports* 12, 1 (2022), 9537.

[2] Ahmed Abbasi, Abdul Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, and Zunera Jalil. 2022. Deep learning for religious and continent-based toxic content detection and classification. *Scientific Reports* 12, 1 (2022), 17478.

[3] Minale A Abebe, Joe Tekli, Fekade Getahun, Richard Chbeir, and Gilbert Tekli. 2017. Overview of event-based collective knowledge management in multimedia digital ecosystems. In *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 40–49.

[4] Usman Ahmed, Rutvij H Jhaveri, Gautam Srivastava, and Jerry Chun-Wei Lin. 2022. Explainable Deep Attention Active Learning for Sentimental Analytics of Mental Disorder. *Transactions on Asian and Low-Resource Language Information Processing* (2022).

[5] Muhammad Waseem Akram, Muhammad Salman, Muhammad Farrukh Bashir, Syed Muhammad Saad Salman, Thippa Reddy Gadekallu, and Abdul Rehman Javed. 2022. A Novel Deep Auto-Encoder Based Linguistics Clustering Model for Social Text. *Transactions on Asian and Low-Resource Language Information Processing* (2022).

[6] Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, et al. 2021. English-Twi Parallel Corpus for Machine Translation. *arXiv preprint arXiv:2103.15625* (2021).

[7] Debajyoty Banik, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Machine learning based optimized pruning approach for decoding in statistical machine translation. *IEEE Access* 7 (2018), 1736–1751.

[8] Debajyoty Banik, Asif Ekbal, Pushpak Bhattacharyya, and Siddhartha Bhattacharyya. 2019. Assembling translations from multi-engine machine translation outputs. *Applied Soft Computing* 78 (2019), 230–239.

[9] Asma Belhadi, Youcef Djenouri, Gautam Srivastava, and Jerry Chun-Wei Lin. 2023. Fast and Accurate Framework for Ontology Matching in Web of Things. *ACM Transactions on Asian and Low-Resource Language Information Processing* (2023).

[10] LU Bin, K Tsou Benjamin, Tao Jiang, Oi Yee Kwong, and Jingbo Zhu. 2010. Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

[11] Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 65–72.

[12] Ondřej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Straňák, Vít Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (26-31), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Reykjavik, Iceland.

[13] Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics* (Berkeley, California) *(ACL '91)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 169–176. https://doi.org/10.3115/981344.981366

[14] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics* 19, 2 (1993), 263–311.

[15] Juryong Cheon and Youngjoong Ko. 2021. Parallel sentence extraction to improve cross-language information retrieval from Wikipedia. *Journal of Information Science* 47, 2 (2021), 281–293.

[16] Anna Dillon, Geraldine Chell, Jase Moussa-Inaty, Kay Gallagher, and Ian Grey. 2021. English Medium Instruction and the potential of translanguaging practices in higher education. *Translation and Translanguaging in Multilingual Contexts* 7, 2 (2021), 153–176.

[17] Martín Fuchs and Paz González. 2022. Perfect-Perfective Variation across Spanish Dialects: A Parallel-Corpus Study. *Languages* 7, 3 (2022), 166.

[18] William A Gale and Kenneth W Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics* 19, 1 (1993), 75–102.

[19] Taher M Ghazal, Mohammad Kamrul Hasan, Siti Norul Huda Abdallah, and Khairul Azmi Abubakkar. 2022. Secure IoMT pattern recognition and exploitation for multimedia information processing using private blockchain and fuzzy logic. *Transactions on Asian and Low-Resource Language Information Processing* (2022).

[20] Mandy Guo, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. *arXiv preprint arXiv:1906.08401* (2019).

[21] Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 187–197.

[22] Abdul Rehman Javed, Saif Ur Rehman, Mohib Ullah Khan, Mamoun Alazab, and Habib Ullah Khan. 2021. Betalogger: Smartphone sensor-based side-channel attack detection and text inference using language modeling and dense multilayer neural network. *Transactions on Asian and Low-Resource Language Information Processing* 20, 5 (2021), 1–17.

[23] Girish Nath Jha. 2010. The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI).. In *LREC*.

[24] Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, Vol. 1. IEEE, 181–184.

[25] Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, Vol. 5. 79–86.

[26] Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 48–54.

[27] Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The IIT Bombay English-Hindi Parallel Corpus. *arXiv preprint arXiv:1710.02855* (2017).

[28] Ivana Kvapilıková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2021. Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining. *arXiv preprint arXiv:2105.10419* (2021).

[29] Guokun Lai, Zihang Dai, and Yiming Yang. 2020. Unsupervised Parallel Corpus Mining on Web Data. *arXiv preprint arXiv:2009.08595* (2020).

[30] Bert Le Bruyn, Martín Fuchs, Martijn van der Klis, Jianan Liu, Chou Mo, Jos Tellings, and Henriëtte De Swart. 2022. Parallel corpus research and target language representativeness: The contrastive, typological, and translation mining traditions. *Languages* 7, 3 (2022), 176.

[31] Lalita Lowphansirikul, Charin Polpanumas, Attapol T Rutherford, and Sarana Nutanong. 2022. A large english–thai parallel corpus from the web and machine-generated text. *Language Resources and Evaluation* 56, 2 (2022), 477–499.

[32] Bin Lu, Benjamin K Tsou, Tao Jiang, Oi Yee Kwong, and Jingbo Zhu. 2010. Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT. In *Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing*. 79–86.

[33] Robert Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. *Machine Translation: From Research to Real Users* (2002), 135–144.

[34] Toshiaki Nakazawa, Shohei Higashiyama, Chenchen Ding, A Kunchukuttan, P Pa, W, I Goto, H. Mino, K Sudoh, and S Kurohashi. 2018. Overview of the 5th Workshop on Asian Translation. In *Proceedings of the 5th Workshop on Asian Translation (WAT2018)*.

[35] David D Palmer and Marti A Hearst. 1997. Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics* 23, 2 (1997), 241–267.

[36] Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 527–534.

[37] Jeffrey C Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the fifth conference on Applied natural language processing*. Association for Computational Linguistics, 16–19.

[38] Michael D Riley. 1989. Some applications of tree-based modelling to speech and language. In *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 339–352.

[39] Manish Shrivastava and Pushpak Bhattacharyya. 2008. Hindi pos tagger using naive stemming: Harnessing morphological information without extensive linguistic knowledge. In *International Conference on NLP (ICON08), Pune, India*.

[40] Jason R Smith, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 403–411.

[41] Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel Corpus for Indigenous Language Translation: Spanish-Mazatec and Spanish-Mixtec. *arXiv preprint arXiv:2305.17404* (2023).

[42] Masao Utiyama and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 72–79.

[43] Masao Utiyama and Hitoshi Isahara. 2007. A Japanese-English patent parallel corpus. *MT summit XI* (2007), 475–482.

[44] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4* 292 (2007), 247.

[45] Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. *arXiv preprint arXiv:2005.06166* (2020).