Neural Machine Translation for Low Resource Languages using Bilingual Lexicon Induced by Comparable Corpora

Sree Harsha Ramesh
Department of Computer Science
University of Massachusetts Amherst
shramesh@cs.umass.edu

Krishna Prasad Sankaranarayanan
Department of Computer Science
University of Massachusetts Amherst
ksankaranara@cs.umass.edu

Abstract

Automatically extracting parallel sentence pairs from the multilingual articles available on the Internet can address the data sparsity problem in building multilingual natural language processing applications, especially in machine translation. In this project, we have used an end-to-end siamese bidirectional recurrent neural network to generate parallel sentences from comparable multilingual articles in Wikipedia. Subsequently, we found that using the enriched dataset showed improved BLEU scores on both NMT and phrase-based SMT systems when compared to training exclusively on carefully curated bilingual corpora.

1 Introduction

Both neural and statistical machine translation approaches are data-hungry and are known to perform poorly in low resource settings. Recent crowd-sourcing efforts and workshops on machine translation have resulted in small amounts of parallel texts for building viable machine translation systems for low-resource pairs (Post et al., 2012). But, they have been shown to suffer from low accuracy (incorrect translation) and low coverage (high out-of-vocabulary rates), due to insufficient training data. In this project, we try to address the high OOV rates in low-resource machine translation systems by leveraging the increasing amount of multilingual content available on the Internet for enriching the bilingual lexicon.

Comparable corpora such as Wikipedia, are collections of topic-aligned but non-sentence-aligned multilingual documents which are rich resources for extracting parallel sentences from. For e.g, Figure 1 shows that there are equivalent sentences on the page about Donald Trump in Tamil and English, and the phrase alignment for an example sentence is shown in Table 2.

Language	# Bilingual Wiki articles	# Curated X-En sentence pairs
Urdu	124,078	35,916
Hindi	121,234	1,495,854
Tamil	113,197	169,871
Telugu	67,508	46,264
Bengali	52,518	23,610
Malayalam	52,224	33,248

Table 1: No. of bilingual articles in Wikipedia against the no. of parallel sentences in the largest X-En corpora available.

Table 1 shows that parallel sentences could be mined from the large number of bilingual articles on Wikipedia to address the scarcity of parallel sentences in the largest reported bilingual corpora for X-En [1]. The illustrated data sparsity can thus be addressed by inducing the scarce bilingual lexicon with sentence-pairs automatically extracted from Wikipedia and also improve the performance of statistical machine translation systems (Irvine and Callison-Burch, 2013). We would be comparing the BLEU scores on SMT and NMT systems with and without the use of the extracted sentence pairs to justify the effectiveness of this method.

Compared to previous approaches which require specialized metadata from document structure or significant amount of hand-engineered features, the neural model for extracting parallel sentences is learned end-to-end using only a small bootstrap set of parallel sentence pairs.

[1] Tamil-En : http://ufal.mff.cuni.cz/~ramasamy/parallel/html/
Hindi-En: http://www.cfilt.iitb.ac.in/iitb_parallel/
Others-En: https://github.com/joshua-decoder/indian-parallel-corpora

Figure 1: A side-by-side comparison of nearly parallel sentences from bilingual Wikipedia articles about Donald Trump in English and Tamil.

English Word	Tamil Word
Donald Trump	டோனால்ட் டிரம்ப்
President	அரசுத்தலைவர்
United States	ஐக்கிய அமெரிக்கா
Casinos	சூதாட்ட விடுதிகள்
Hotels	தங்கும் விடுதிகள்

Table 2: Phrase-aligned EN-TA pairs from Fig 1

2 Related Work

A lot of work has been done on the problem of sentence alignment from comparable corpora, but most of them (AbdulI-Rauf and Schwenk, 2009; Irvine and Callison-Burch, 2013; Yasuda and Sumita, 2008) use a pre-existing translation system as a precursor to ranking the candidate sentence pairs, which the low resource language pairs are not at the luxury of having. There are statistical machine learning approaches as described in (Munteanu and Marcu, 2005), where a MaxEnt classifier is used that relies on surface level features such as word overlap in order to obtain parallel sentence pairs. However, the deep neural network model used in this paper is probably the first of its kind, which does not need any feature engineering and also does not need a pre-existing translation system.

An example of using comparable corpora to improve translation of low resource languages successfully, was described in (Irvine and Callison-Burch, 2013) where they used comparable corpora in order to overcome the failure of statistical machine translation systems on low resource languages due to low accuracy and low coverage. Bilingual lexicon induction was used to train a discriminative classifier which makes predictions on test set words paired with all possible translations. The translation accuracy was in turn improved by using comparable corpora in which additional features were estimated in the phrase tables and used during tuning and decoding. The out of vocabulary (OOV) words formed an integral part of this method to compute the gains in the BLEU evaluation metric. The usage of comparable corpora by adding the top K- translations resulted in an appreciable improvement in accuracy when both features and translations are estimated over phrase pairs than when estimated on the OOV words. However unlike this work, our initial phrase tables were estimated from extremely limited parallel corpora for low resource languages but also translation pairs from comparable corpora.

(Munteanu and Marcu, 2005) proposed a parallel sentence extraction system which used comparable corpora from newspaper articles to extract the parallel sentence pairs. In this procedure, a maximum entropy classifier is designed for all sentence pairs possible from the Cartesian product of a pair of documents and passed through a sentence-length ratio filter in order to obtain candidate sentence pairs. SMT systems were trained on the extracted sentence pairs using the additional features from the comparable corpora like distortion and position of current and previously aligned sentences. This resulted in a state of the art approach with respect to the translation performance of low resource languages.

Similar to our proposed approach, (Barrón-Cedeño et al., 2015) showed how using parallel documents from Wikipedia for domain specific alignment would improve translation quality of SMT systems on in-domain data. In this method, similarity between all pairs of cross-language sentences with different text similarity measures are estimated. The issue of domain definition is overcome by the use of IR techniques which use the characteristic vocabulary of the domain to query a Lucene search engine over the entire corpus. The candidate sentences are defined based on word overlap and the decision whether a sentence pair is parallel or not using the maximum entropy classifier. The difference in the BLEU scores between out of domain and domain-specific translation is proved clearly using the word embeddings from characteristic vocabulary extracted using the extracted additional bitexts.

(AbduI-Rauf and Schwenk, 2009) make this process of extracting parallel sentences easier without the use of a classifier. Instead, target language candidate sentences are found using the translation of source side comparable corpora. Sentence tail removal is a means by which removing the tail parts of sentence pairs which differ only at the end. This along with parallel sentences enhanced the BLEU score and helped to determine if the translated source sentence and candidate target sentence are parallel by measuring the word and translation error rate. This method succeeds in eliminating the need for domain specific text by using the target side as a source of candidate sentences. However, this approach is not feasible if there isn't a good source side translation system to begin with, like in our case.

Yet another approach which uses an existing translation system to extract parallel sentences from comparable documents was proposed by (Yasuda and Sumita, 2008). They describe a framework for machine translation using multilingual Wikipedia articles. The parallel corpus is assembled by comparing the sentence-level MT evaluation scores between Japanese and English sentences with their sentence counterparts obtained via translation. The sentence aligned corpus is finally created by aligning the sentences from the above method. After filtering out the unaligned pairs, the effectiveness of the MT evaluation measure is calculated as the performance of the system before and after the filtering. The automatic update of the lexicon using the aligned sentences happens by the elemental technology of STM unlike previous approaches. The results showed a marked increase in translation accuracy in turn maintaining high alignment quality.

3 Approach

In this section, we would be describing the entire pipeline involved in training a parallel system extraction system, and using that to infer and decode high precision nearly parallel sentences from parallel article pages collected from Wikipedia. The same has been depicted in Figure 2.

3.1 Bootstrap Dataset

The parallel sentence extraction system needs a sentence aligned corpus which has been carefully curated. These sentences were used as the ground truth pairs when we trained the model to classify parallel sentence pair from non-parallel pairs.

3.2 Negative Sampling

The binary classifier described in the next section, assigns a translation probability score to a given sentence pair, after learning from examples of translations and negative examples of non-translation pairs. For, this we make a simplistic assumption that the parallel sentence pairs found in the bootstrap dataset are unique combinations, which fail being translations of each other, when we randomly pick a sentence from both the sets. Thus, there might be cases of false negatives in due to the reliance on unsupervised random sampling for generation of negative labels.

Therefore at the beginning of every epoch, we randomly sample m negative sentences of the target

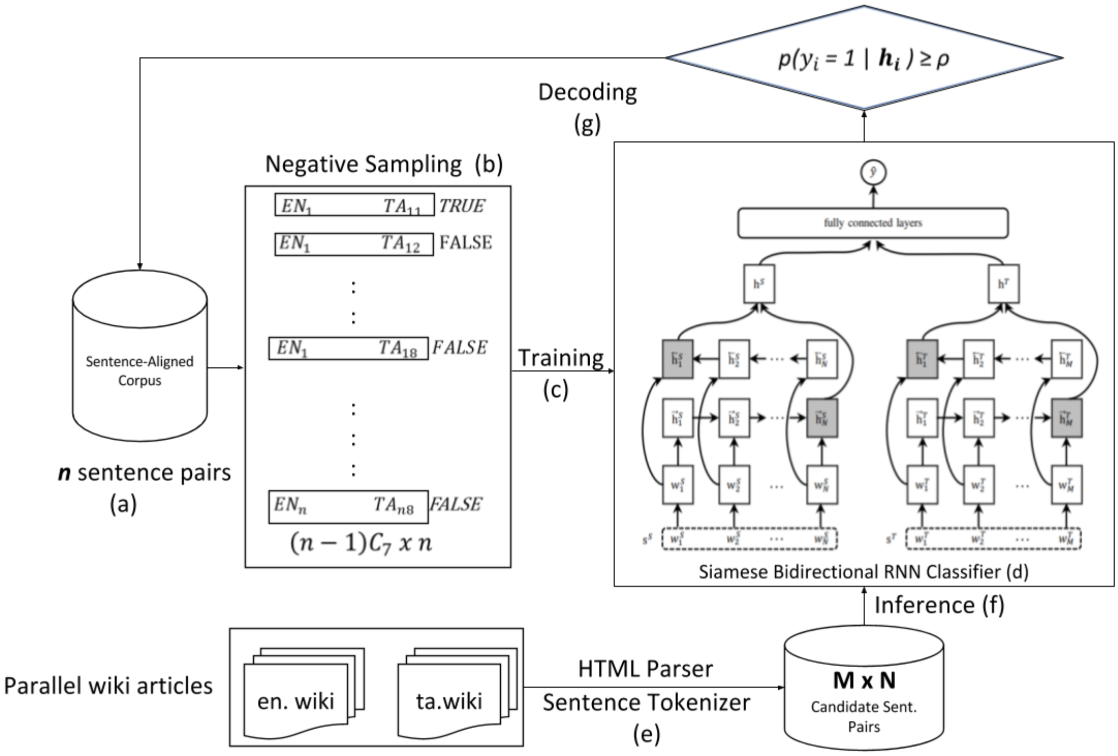


Figure 2: Architecture for the parallel sentence extraction system showing training and inference pipelines. EN - English, TA - Tamil

language for every source sentence. From a few experiments and also from the literature, we converged on m=7, to be performing the best given our compute constraints.

3.3 Model

Here, we describe the neural network architecture as shown in (Grégoire and Langlais, 2017), where the network learns to estimate the probability that the sentences in a given sentence pair, are translations of each other, $p(y_i = 1|\mathbf{s}_i^S, \mathbf{s}_i^T)$, where $\boldsymbol{s}_i^S$ is the candidate source sentence in the given pair, and $\boldsymbol{s}_i^S$ is the candidate target sentence.

3.3.1 Training

As illustrated in Figure 2 (d), the architecture uses a siamese network (Bromley et al., 1994) consisting of a bidirectional RNN (Schuster and Paliwal, 1997) sentence encoder with recurrent activation functions such as long short-term memory units (LSTM) (Hochreiter and Schmidhuber, 1997) or gated recurrent units (GRU) (Cho et al., 2014) learning a vector representation for the source and target sentences and the probability of any given pair of sentences being translations of each other. For seq2seq architectures, especially in translation, we have found the that the recommended activation unit is GRU, and all our experiments use this over LSTM.

The forward RNN reads the variable-length sentence and updates its recurrent state from the first token until the last one to create a fixed-size continuous vector representation of the sentence. The backward RNN processes the sentence in reverse. In our experiments, we use the concatenation of the last recurrent state in both directions as a final representation $\mathbf{h}_i^S = [\overrightarrow{\mathbf{h}}_{i,N}^S ; \overleftarrow{\mathbf{h}}_{i,1}^S]$

$$\mathbf{w}_{i,t}^S = \mathbf{E}^{S^\top} \mathbf{w}_k \tag{1}$$

$$\overrightarrow{\mathbf{h}}_{i,t}^S = \phi(\overrightarrow{\mathbf{h}}_{i,t-1}^S, \mathbf{w}_{i,t}^S) \tag{2}$$

$$\overleftarrow{\mathbf{h}}_{i,t}^S = \phi(\overleftarrow{\mathbf{h}}_{i,t+1}^S, \mathbf{w}_{i,t}^S) \tag{3}$$

ϕ can be any recurrent activation function, such as LSTM or GRU. After both source and target sentences have been encoded, we capture their matching information by using their element-wise product and absolute element-wise difference. We estimate the probability that the sentences are translations of each other by feeding the matching vectors

Extracted Tamil Sentences from Wikipedia	Model generated Parallel English Sentences From Wikipedia	Translation of Extracted Tamil Sentence for Comparison (Google Translate)	Translation Probability
சிலி , எதியோப்பியா பிஜி , இலங்கை , மற்றும் உஸ்பெகிஸ்தான் ஆகியன தமது முதல் பராலிம்பிக் பதக்கங்களை வென்றன.	Athletes from Chile , Ethiopia , Fiji , Sri Lanka , and Uzbekistan won their first Paralympic medals.	Chile, Ethiopia Fiji, Sri Lanka, and Uzbekistan won their first paralympic medals.	0.991
இந்தியப் பிரதமர் சவகர்லால் நேரு சனவரி 16, 1955இல் புதுச்சேரிக்கு வருகை புரிந்தார்.	Prime Minister of India Jawaharlal Nehru visited Pondicherry on 16 January 1955.	Indian Prime Minister Shivakral Nehru visited Puducherry on January 16, 1955.	0.994
ஜிம்மி டேலி Jimmy Daley , பிறப்பு: செப்டம்பர் 24 1973), இங்கிலாந்து அணியின் துடுப்பாட்டக்காரரர்.	Jimmy Daley (born 24 September 1973) is a retired English cricketer .	Jimmy Daley (born September 24, 1973) is an English cricketer.	0.993

Table 3: A sample of parallel sentences extracted from Wiki En-Ta articles. The translation of the extracted Tamil sentence in English is also provided.

into fully connected layers:

$$\mathbf{h}_i^{(1)} = \mathbf{h}_i^S \odot \mathbf{h}_i^T \tag{4}$$

$$\mathbf{h}_i^{(2)} = |\mathbf{h}_i^S - \mathbf{h}_i^T| \tag{5}$$

$$\mathbf{h}_i = tanh(\mathbf{W}^{(1)}\mathbf{h}_i^{(1)} + \mathbf{W}^{(2)}\mathbf{h}_i^{(2)} + \mathbf{b}) \tag{6}$$

$$p(y_i = 1|\mathbf{h}_i) = \sigma(\mathbf{W}^{(3)}\mathbf{h}_i + c) \tag{7}$$

where σ is the sigmoid function, $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{W}^{(3)}$, $\mathbf{b}$ and c are model parameters. The model is trained by minimizing the cross entropy of our labeled sentence pairs:

$$\begin{aligned}\mathcal{L} = -\sum_{i=1}^{n(1+m)} & y_i \log \sigma(\mathbf{W}^{(3)}\mathbf{h}_i + c) \\ & - (1 - y_i)\log(1 - \sigma(\mathbf{W}^{(3)}\mathbf{h}_i + c))\end{aligned} \tag{8}$$

3.3.2 Inference

For prediction, a sentence pair is classified as parallel if the probability score is greater than or equal to a decision threshold ρ that we need to fix. We found that to get high precision sentence pairs, we had to use $\rho = 0.99$, and if we were able to sacrifice some precision for recall, a lower $\rho = 0.80$ of 0.80 would work in the favor of reducing OOV rates.

$$\hat{y}_i = \begin{cases} 1 & \text{if } p(y_i = 1|\mathbf{h}_i) \geq \rho \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

4 Experiments

4.1 Dataset

We experimented with two language pairs - English - Hindi and English - Tamil. We first trained parallel sentence extraction systems for both Tamil-English and Hindi-English using the above-mentioned deep architecture and the following bootstrap set of parallel corpora:

- An English-Tamil parallel corpus (Ramasamy et al., 2014) containing a total of $169,871$ sentence pairs, composed of $3,984,038$ English Tokens and $2,776,397$ Tamil Tokens.

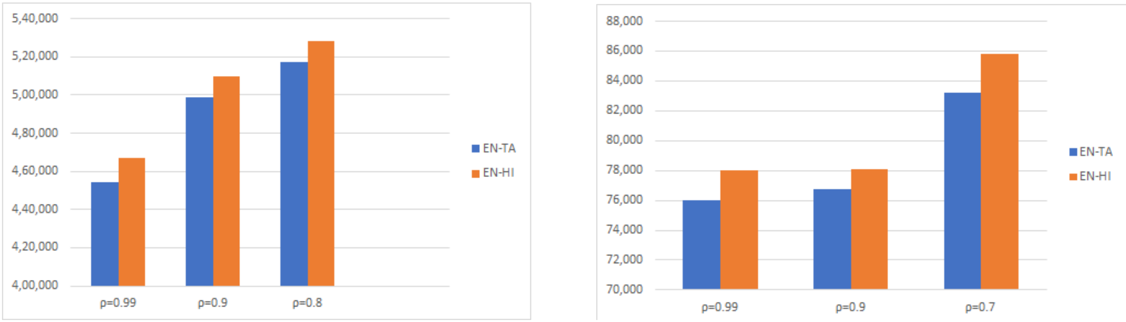
- An English-Hindi parallel corpus (Kunchukuttan et al., 2017)containing a total of $1,492,827$ sentence pairs.

Subsequently, we extracted parallel sentences using the trained model, and parallel articles collected from Wikipedia. [2]. There were 67,449 parallel Tamil-English and 58,802 Hindi-English titles on the latest Wikimedia dumps.

4.2 Evaluation Metrics

For the evaluation of the performance of our sentence extraction models, we looked at a few sentences manually, and have done a qualitative analysis, as there was no ground truth sentences for sentences extracted from Wikipedia. In the table 3, we can see the qualitative accuracy for some parallel sentences extracted from Tamil. The sentences extracted from Tamil, have been translated to English using Google Translate, so as to facilitate a comparison with the sentences extracted from English.

[2]Tamil: dumps.wikimedia.org/tawiki/latest/
Hindi: dumps.wikimedia.org/hiwiki/latest/



(a) Without greedy decoding

(b) With greedy decoding

Figure 3: No of parallel sentences extracted from 10,000 parallel Wikipedia article pairs using different thresholds and decoding methods

For the statistical machine translation and neural machine evaluation we use the BLEU score (Papineni et al., 2002) as an evaluation metric using the multi-bleu script from Moses (Koehn et al., 2007).

4.3 Sentence Alignment

Figures 3a shows the no of high precision sentences that were extracted at $\rho = 0.99$ without greedy decoding. So, a sentence that's been already extracted, is not precluded from being considered again. Hence, the number of sentences are of an order of magnitude higher than with decoding, which has been shown in Figure 3b.
For training the machine translation systems, we used high precision sentences extracted with greedy decoding.

4.4 Machine Translation

4.4.1 SMT

We trained Phrase Based SMT systems with Moses (Koehn et al., 2007). We used the grow-diag-final-and heuristic for extracting phrases, lexicalised reordering and Batch MIRA (Cherry and Foster, 2012) for tuning (the default parameters on Moses). We trained 5-gram language models with Kneser-Ney smoothing using KenLM (Heafield et al., 2013). With these parameters, we trained systems for En-Ta and En-Hi language pairs, with and without the use of parallel sentence pairs.

4.4.2 NMT

For training neural machine translation models, we used the TensorFlow (Abadi et al., 2016) implementation of OpenNMT (Klein et al.) with the state-of-the-art attention-based transformer architecture (Vaswani et al., 2017) for neural translation. The figures for the NMT models were better than for SMT for both En-Ta and En-Hi pairs, as can be seen in Table 4.

5 Conclusion

In this paper, we evaluated the benefits of using a neural network procedure to extract parallel sentences. Unlike traditional translation systems which make use of multi-step classification procedures, this method requires just a parallel corpus to extract parallel sentence pairs using a Siamese BiRNN encoder using GRU as the activation function.

This method is extremely beneficial for translating language pairs with very little parallel corpora. These parallel sentences facilitate significant improvement in machine translation quality when compared to a generic system as has been shown in our results. Since this is a model trained on negative sampling, it allowed us to use a higher threshold of 0.99 which greatly improved our performance, in terms of quality of translation.

The experiments are shown for English-Tamil and English-Hindi language pairs. Our model achieved a marked percentage increase in the BLEU score for both EN-TA and En-HI language pairs.The percentage increase in BLEU scores was 11.03% and 14.7% for EN-TA and EN-HI pairs respectively.

Currently, this approach does not handle out-of-vocabulary words. Our methodology has huge potential forr future improvement and analysis. One of the main features to look out for improvement would be how the approach would apply in non domain specific language pairs. Another major area of utility in future would be in other domains of NLP such as speech transcription, named entity recognition and summarization.

Training Data	Model	BLEU	#Sents
IIT Bombay En-Hi	SMT	2.96	200,000
+ Wiki Extracted =0.99	SMT	3.57(+0.61)	+77,988
IIT Bombay En-Hi	NMT	3.46	200,000
+ Wiki Extracted =0.99	NMT	3.97(+0.51)	+77,988
Ramasamy et.al En-Ta	SMT	4.02	169,871
+ Wiki Extracted =0.99	SMT	4.57(+0.55)	+75,970
Ramasamy et.al En-Ta	NMT	4.53	169,871
+ Wiki Extracted =0.99	NMT	5.03(+0.50)	+75,970

Table 4: BLEU score results for En-Hi and En-Ta

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Sadaf AbduI-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve smt performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 16–23. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Cristina España Bonet, Josu Boldoba Trapote, and Luís Márquez Villodre. 2015. A factory of comparable corpora from wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13. Association for Computational Linguistics.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Francis Grégoire and Philippe Langlais. 2017. A deep neural network approach to parallel sentence extraction. *arXiv preprint arXiv:1709.09783*.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 2013. Scalable modified kneser-ney language model estimation. In *ACL (2)*, pages 690–696.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ann Irvine and Chris Callison-Burch. 2013. Combining bilingual and comparable corpora for low resource machine translation. In *WMT@ ACL*, pages 262–270.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2017. The iit bombay english-hindi parallel corpus. *arXiv preprint arXiv:1710.02855*.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

Loganathan Ramasamy, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2014. Entam: An english-tamil parallel corpus (entam v2. 0).

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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

Keiji Yasuda and Eiichiro Sumita. 2008. Method for building sentence-aligned corpus from wikipedia. In *2008 AAAI Workshop on Wikipedia and Artificial Intelligence (WikiAI08)*, pages 263–268.