MUCS@Adap-MT 2020: Low Resource Domain Adaptation for Indic Machine Translation

Asha Hegde
Department of Computer Science
Mangalore University
hegdekasha@gmail.com

H. L. Shashirekha
Department of Computer Science
Mangalore University
hlsrekha@gmail.com

Abstract

Machine Translation (MT) is the task of automatically converting the text in source language to text in target language by preserving the meaning. MT task usually require large corpus for training the translation models. Due to scarcity of resources very less attention is given to translating into low resource languages and in particular into Indic languages. In this direction, a shared task called "Adap-MT 2020: Low Resource Domain Adaptation for Indic Machine Translation" is organized to illustrate the capability of general domain MT when translating into Indic languages and low resource domain adaptation of MT systems. In this paper, we, team MUCS, describe a simple word extraction based domain adaptation approach applied to English-Hindi MT only. MT in the proposed model is carried out using Open-NMT - a popular Neural Machine Translation tool. A general domain corpus is built effectively combining the available English-Hindi corpora and removing the duplicate sentences. Further, domain specific corpora is updated by extracting the sentences from generic corpus that match with the vocabulary of the domain specific corpus. The proposed model is exhibited satisfactory results for small domain specific AI and CHE corpora in terms of Bilingual Evaluation Understudy (BLEU) score with 1.25 and 2.72 respectively. Further, this methodology is quite generic and can easily be extended to other low resource language pairs as well.

1 Introduction

Machine Translation (MT) acts as a bridge for cross-language communication in Natural Language Processing (NLP). It handles perplexity problems between two languages while preserving its meaning. MT was one of the initial tasks taken up by computer scientists and the research in this field is going on for last 50 years. MT task was initially handled with dictionary matching techniques and slowly upgraded to rule-based approaches (Dove et al., 2012). To resolve knowledge acquisition issues corpus based approaches became popular and bilingual parallel corpora was used to acquire translation knowledge (Britz et al., 2017). Along with corpus based approaches, hybrid MT approaches also became popular as these approaches promise state-of-the-art result.

The recent shift to large-scale analytical techniques has resulted in very significant improvements in the quality of MT. Neural Machine Translation (NMT) - a corpus based approach has gained attention of the MT researchers. NMT is the task of translating text from one natural language (source) to another natural language (target) using most commonly, Recurrent Neural Networks (RNN), specifically the Encoder-Decoder or Sequence-to Sequence models (Sutskever et al., 2014). Further, unlike conventional translation systems, all parts of the neural translation model are trained jointly (end-to-end) to maximize the translation performance (Bahdanau et al., 2014). In an NMT system, a bidirectional RNN, known as encoder is used by the Neural Network (NN) to encode a source sentence for a second RNN, known as decoder which is used to predict words in the target language. This encoder-decoder architecture can be designed with multiple layers to increase the efficiency of the system. Now, NMT has become an effective alternative to traditional Phrase-Based Statistical Machine Translation (Patil and Davies, 2014).

1.1 Challenges of NMT

In spite of its popularity, NMT faces the following challenges

- Normally NMT require a large dataset for training the model and powerful computa-

Proceedings of the 17th International Conference on Natural Language Processing: Adap-MT 2020 Shared Task, pages 24–28
Patna, India, December 18 - 21, 2020.

tional resource to build NN with sufficient amount of hidden layers.

- NMT is inconsistent in handling rare words. Since these words are sparsely available in the network, learning and inferencing them is not efficient.

- Though many experiments are being carried out to handle long sentences, long term dependency issue is still considered as a major problem in NMT (Tien and Minh, 2019).

The main objective of this work is to investigate efficient strategies to perform English to Hindi MT using sufficient amount of general domain corpora and very small domain specific corpora. Rest of the paper is structured as follows: Section 2 gives the brief description about domain adaptation and different approaches to domain adaptation followed by the methodology in Section 3. Experiments and results are given in Section 4 and conclusion in Section 5.

2 Domain Adaptation for NMT

Dataset plays a crucial role in NN based translation models. Huge amount of quality dataset for training results in good translation performance whereas small dataset results in poor translation performance. Hence, if the dataset is small, effective management of such dataset for NN based translation will be the key for better translation performance. Domain adaptation techniques that transfer existing knowledge to new domains as much as possible is one method in this direction. Domain Adaptation (DA) is a sub-discipline of machine learning in which a model trained on a source distribution is used in the context of a different (but related) target distribution. In simple words, it is the ability to apply an algorithm trained in one domain to a different domain or updating one corpus using another corpus.

While the big generic corpus will help to avoid out-of-vocabulary problem and unidiomatic translations, the small specialized corpus will help to capture terminology and vocabulary that is required for the translation (Šoštarić et al., 2019). Few effective DA approaches which promise better translation performance are as follows:

- Incremental Training and Re-training - In this approach, initially a model is trained on a huge generic corpus and then the same model is re-trained on a small domain specific corpus. This approach has two phases: i) pre-processing and training of huge generic corpus and ii) pre-processing the new domain specific corpus and re-training the base model on the domain specific corpus (Kalimuthu et al., 2019).

- Ensemble of decoding - In this approach, the base model is trained on generic dataset and the model is re-trained on domain specific dataset. Then instead of combining dataset, both the models are combined during translation (Chu and Wang, 2018).

- Combining Training Data - This approach is a simple and effective DA approach compared to all other approaches. In this approach, both the corpora are combined and this new corpus is used for training ie., huge generic corpus is combined with domain specific corpus and then this new corpus is used for training (Chu and Wang, 2018).

- Data Augmentation - In this approach, size of the domain specific dataset is increased using phrase based translation technique. The information related to word alignment is extracted from the corpus and then this information is used to build n-gram model to construct new dataset. Further, duplicates are discarded to avoid redundancy (Xia et al., 2019).

Table 1: Details of General domain English-Hindi parallel corpus

Resource	**No. of parallel sentences**	**No. of words**
IIT Bombay	2,00,000	6,28,56,567
Bible	62,073	4,10,589
globalvoices	2,299	1,70,116
CVIT-MKB	5,272	3,54,128

3 Methodology

Despite the considerable advances in MT models, translation of low-resource languages is still an unresolved issue and DA approaches are promising considerable performance in this direction. In the proposed work, a DA approach of combining both generic dataset and domain specific dataset based on the vocabulary of domain specific dataset is

used to conduct effective training and inference for translation using openNMT- a popular open source tool (Klein et al., 2018). OpenNMT accepts only primarily cleaned dataset as its input. Therefore, noise such as initial space, end space, blank lines and special characters have been removed from the bilingual parallel corpus. This pre-processing is carried out for both generic corpus and domain specific corpora. Then vocabulary of the domain specific corpora is constructed and sentences that contain any of the words in this vocabulary are extracted from the generic corpus. Finally, these extracted sentences are added to the domain specific corpus and the updated corpus is used to train the translation model. Table 2 illustrates a sample sentence from generic corpus and from domain specific corpus along with their vocabulary. The word 'queen' which is present in domain specific corpus is also present in the generic corpus. Hence, that sentence from the generic corpus will be extracted and added to the domain specific corpus.

3.1 Dataset

Dataset and the preparation of dataset for training the translation model play a major role in MT. This data preparation process is carried out at different levels to conduct effective translation.

General domain English-Hindi corpus is constructed by combining various open source corpora namely English-Hindi parallel corpus open sourced by IIT Bombay[1], English-Hindi bible corpus[2], Globalvoices[3] and CVIT-MKB[4]. Then this newly constructed generic corpus is pre-processed so that the corpus can be used to train in openNMT. Sufficient training and validation dataset is used which is the basic requirement of openNMT.

AI English-Hindi corpus is pre-processed and combined with general domain English-Hindi corpus based on the vocabulary of AI English-Hindi corpus. Then this new corpus is used for translation in openNMT model.

Chemistry English-Hindi corpus is pre-processed and combined with general domain English-Hindi corpus based on the vocabulary of CHE English-Hindi corpus. Then this new corpus is used for translation in openNMT model.

Details of general domain English-Hindi parallel corpus are shown Table 1 and details of AI and CHE corpora are shown in Table 4. Table 3 shows the details of domain specific dataset after applying DA and details of train and validation dataset used for the experiments are shown in Table 6.

4 Experimental setup

English to Hindi MT is implemented using openNMT which is considered as the most sophisticated generalized translation tool that provides easy modifications. As this model requires GPU, we set up this experiment in Google colaboratory. Translation experiments are carried out by continuous tuning of the model to conduct better training. Initially, this model is trained using a huge generic corpus then the same set up is used for domain specific corpus. As the given domain specific corpora are very small to conduct efficient translation, training data of domain specific corpora is combined with generic corpus based on vocabulary of the domain specific corpora and the training is continued with the same set up.

4.1 Result

The proposed model predicts Hindi sentences for the given English test sentences and the sample snapshot of the model is shown in Figure 1 and the performance measure of the proposed model in terms of accuracy and perplexity is shown in Table 5. Further, the proposed system is evaluated separately using BLEU score (Papineni et al., 2002) for both generic corpus and domain specific corpora. Though there are many challenges with the test dataset, considerable results are obtained for both generic corpus and domain specific corpora.

4.2 Result Analysis

The results obtained for the given test set with respect to general domain corpus shows 63.43% accuracy with 20.51 perplexity using openNMT model. This model shows considerable accuracy for the generic corpus as it contains lots of challenges related to alignment, mixing of different script, length of the sentences etc. Then, the results obtained for translating the given test set with respect to domain specific AI corpus in the same setup shows 30.63% accuracy with 45.68 perplexity. As this corpus is very small to conduct translation the same is replicated in the result ie., it exhibits poor translation. Then, after applying proposed DA approach the model shows improvement in both accuracy

[1] http://www.cfilt.iitb.ac.in/iitb$_parallel$
[2] http://opus.nlpl.eu/bible-uedin.php
[3] http://opus.nlpl.eu/GlobalVoices.php
[4] http://preon.iiit.ac.in/ jerin/bhasha/

Table 2: Sample sentences

corpus	Sentence	Vocabulary
Generic corpus	The **Queen** said: Know my nobles that a gracious letter has been delivered to me.	**queen**, said, know, nobles, gracious, letter, delivered
Domain specific corpus	Example one, in a bee hive, there are many thousands of workers bee that all serve one **queen** bee	example, one, bee, hive, thousands, workers, serve, **queen**

SENT 16: ['The', 'interactions,', 'reactions', 'and', 'transformations', 'that', 'are', 'studied',

'in', 'chemistry', 'are', 'usually', 'the', 'result', 'of', 'interactions', 'between', 'atoms,', 'leading',

'to', 'rearrangements', 'of', 'the', 'chemical', 'bonds', 'which', 'hold', 'atoms', 'together.']

PRED 16: स्टेबिलिटी और तहपत्हेस के बीच में जो कार्बोहाइड्रेट्स का अध्ययन करते हैं, वे हाइड्रोजन बॉन्ड्स के बीच होते हैं, जो हाइड्रोजन बॉन्ड्स के बीच हाइड्रोजन बॉन्ड्स होते हैं।

PRED SCORE: -45.7755

SENT 17: ['Such', 'behaviors', 'are', 'studied', 'in', 'a', 'chemistry', 'laboratory.']

PRED 17: यह chemistry laboratory. में लिखा हुआ लिखा है।

PRED SCORE: -10.0361

SENT 18: ['The', 'chemistry', 'laboratory', 'stereotypically', 'uses', 'various', 'forms', 'of',

'laboratory', 'glassware.']

PRED 18: जेवनारों में गवैयों का निर्माण करने के लिए पृथक्करण का निर्माण किया जाता है।

PRED SCORE: -26.7975

Figure 1: Predicted English-Hindi sentences using openNMT

Table 3: Details of domain specific English-Hindi parallel corpora after domain adaptation (for training)

Corpus name	No. of parallel sentences	No. of words	Vocab Size
AI	2,28,079	6,66,42,961	98,606
CHE	2,27,873	6,62,58,875	1,00,006

Table 4: Details of domain specific English-Hindi parallel corpora before domain adaptation (for training)

Corpus name	No. of parallel sentences	No. of words
AI	4,383	8,05,483
CHE	3,567	13,72,980

and perplexity ie., 41.98% and 38.52 respectively. Because of DA technique used for translation, domain specific dataset is increased to capture rare words that improves the translation. Further, the results obtained for translating the given test set with respect to domain specific CHE corpus using

Table 5: Performance measurement of the model

Corpus Name	Accuracy	Perplexity
Generic Corpus	63.43	20.51
AI (Before DA)	30.63	45.68
CHE (Before DA)	31.57	40.48
AI (After DA)	41.98	38.52
CHE (After DA)	42.87	29.25

Table 6: Details of training and validation sentences used for the model

Corpus name	**No. of Training sentences**	**No. of validation sentences**
Generic	2,69,400	20,244
AI	2,65,383	20,400
CHE	2,46,867	20,300

openNMT shows 31.57% accuracy with 40.48 perplexity. Then, proposed DA approach is applied and newly constructed corpus is used in the model. It shows improvement in both accuracy and perplexity ie., 42.87% and 29.25 respectively.

5 Conclusion and Future work

In this English-Hindi translation work, a huge generic corpus and small domain specific corpora are used for translation in openNMT. Further, a simple domain adaptation technique is used to tackle translation issues of low-resource languages. As this approach is language independent it can easily be extended to other low-resource languages. Further, these experiments have exhibited satisfactory results for both generic corpus and domain specific corpora.

We would like to explore different preprocessing techniques that helps to translate low resource languages efficiently.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Denny Britz, Anna Goldie, Minh-Thang Luong, and Quoc Le. 2017. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*.

Catherine Dove, Olga Loskutova, and Ruben de la Fuente. 2012. What's your pick: Rbmt, smt or hybrid. In *Proceedings of the tenth conference of the Association for Machine Translation in the Americas (AMTA 2012). San Diego, CA*.

Marimuthu Kalimuthu, Michael Barz, and Daniel Sonntag. 2019. Incremental domain adaptation for neural machine translation in low-resource settings. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 1–10.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M Rush. 2018. Opennmt: Neural machine translation toolkit. *arXiv preprint arXiv:1805.11462*.

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

Sumant Patil and Patrick Davies. 2014. Use of google translate in medical communication: evaluation of accuracy. *Bmj*, 349:g7392.

Margita Šoštarić, Nataša Pavlović, and Filip Boltužić. 2019. Domain adaptation for machine translation involving a low-resource language: Google automl vs. from-scratch nmt systems. *Translating and the Computer*, 41:113–124.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Ha Nguyen Tien and Huyen Nguyen Thi Minh. 2019. Long sentence preprocessing in neural machine translation. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. *arXiv preprint arXiv:1906.03785*.