DesiPayanam: developing an Indic travel partner

Diviya K N, Mrinalini K, Vijayalakshmi P, Thenmozhi J,
Department of Electronics and Communication Engineering,
Sri Sivasubramaniya Nadar College of Engineering, Chennai
Nagarajan T
Department of Computer Science and Engineering,
Shiv Nadar University, Chennai

Abstract

Domain-specific machine translation (MT) systems are essential in bridging the communication gap between people across different businesses, economies,and countries. India, a linguistically rich country with a booming tourism industry is a perfect market for such an MT system. On this note, the current work aims to develop a domain-specific transformer-based MT system for Hindi-to-Tamil translation. In the current work, neural-based MT (NMT) model is trained from scratch and the hyperparameters of the model architecture are modified to analyse its effect on the translation performance. Further, a finetuning approach is adopted to finetune a pretrained transformer MT model to better suit the tourism domain. The proposed experiments are observed to improve the BLEU scores of the translation system by a maximum of 1% and 4% for the training from scratch and finetuned systems respectively.

1 Introduction

India is a multilingual nation whose diversity is reflected in the usage of local languages for communication across the states. It is essential to ensure that knowledge or information transfer between the states occur without flaw for various sectors like education, business, tourism, and health to flourish (Mrinalini and Vijayalakshmi, 2015). On this note, the current work aims to develop a machine translation system to enable translation of domain-specific content between Indic languages. While several available translation engines can be used for this purpose, their performance for domain-specific texts is limited.

Ever since the advent of transformer neural models, it has been widely used for various tasks like machine translation, speech processing, pattern recognition etc (Vaswani et al., 2017). Further, transformer-based NMT models have been found to enable the development of low-resource machine translation systems for low-resource Indic languages (Pal et al., 2023). On this note the current work aims to develop a Hindi-to-Tamil translation system for the travel and tourism domain.

Two sets of experiments are carried out to determine the best approach to build a domain-specific MT system. To begin with, a transformer-based NMT model is trained from scratch and the effect of parameter tuning on the model is analyzed. Secondly, the effect of using domain-specific data to finetune a pretrained model is analyzed. For both of these experiments a large amount of parallel data is required which is discussed further in the following section. Section 3 and 4 discuss the above two experiments and their effects in detail. Section 5 concludes the findings and future direction of the current work.

2 Dataset and data pre-processing

As discussed in the previous section, the current work develops a machine translation system using NMT modelling for which huge amount of training data is required. Towards this end, the current work makes use of the following datasets.

- **Samanantar:** This is a publicly available parallel corpora for Indic languages with over 49.7 million sentence pairs between English and 11 Indic languages [1]. In the current work, the Hindi-Tamil corpus from this dataset is used to train the NMT models. The Hindi-Tamil parallel corpora consists of approximately 25 lakh parallel sentences from different domains

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

Table 1: Model Parameters

Parameter	Baseline	FF-4096	Attn-16	W2V-1048
Transformer_ff	2048	4096	4096	4096
Attention heads	8	8	16	16
Vector size	512	512	512	1048

and sources such as, politics, entertainment, e-learning, tourism, health etc.

- **TDIL data:** As discussed earlier, the current work concentrates on developing a machine translation system for travel and tourism content [2]. Towards this end, domain-specific Hindi-Tamil text data from the Technology Development for Indian Languages (TDIL) is considered. A total of 25,000 Hindi-Tamil parallel sentences pertaining to tourism is available in the corpus.

- **Microsoft data:** This data source contains 40 hours of audio data along with corresponding text transcriptions in Tamil [3]. While, the dataset was developed to support speech research, the current work makes use of the text transcripts to generate a new set of parallel sentences for NMT training. This dataset was chosen with the aim to enable translation that are more conversational in nature. In order the create the parallel dataset, the text transcripts in Tamil are Google translated to Hindi. In this way, a total of 19,000 conversational text sentences are generated in this dataset.

2.1 Data pre-processing

The above discussed corpora are found to have errors and non-essential formatting issues which could affect the quality of the final machine translation model. Thus, in the current work pre-processing techniques are used to improve the quality of the text before model training. To begin with, several duplicate sentence pairs, incomplete translations, and empty translations are eliminated. Following this, non-essential punctuation marks, HTML tags, and other auxiliary symbols are removed.

Additionally, in case of the Samanantar dataset a few inconsistencies were observed in terms of availability of correct translation pairs. Thus, the 25 Lakh Hindi sentences were translated to Tamil using Google translate to create the final Samanantar dataset used to train the MT models.

2.2 Test dataset

Two datasets are used to evaluate the performance of the MT systems. The first set consists of 918 sentences which are user-formulated. Google forms were sent to Hindi and/or Tamil speaking users asking them to suggest relevant questions, answers, statements, and information related to the travel domain. The reference sentences for the 918 sentences were generated using Google translate.

For the second set of test data, 2000 parallel sentences pertaining to travel and tourism were identified in the Samanantar dataset and were kept aside for testing. The performance of the translation systems in the current work are evaluated using the language modelling-based BLEU score (Papineni et al., 2002), and the vector-based BERTScore (Zhang et al., 2019) and SBSim score (Mrinalini et al., 2022).

3 Model training and finetuning model parameters

Machine translation systems can be developed using different techniques such as rule-based, statistical-based and neural-based approaches. The advantages and disadvantages of these approaches have been debated over time and it has been concluded that upon the availability of huge amount of training data and computation power neural-based approaches to modelling have an upper hand. On this note, in the current work an NMT model is trained using the OpenNMT toolkit [4] to enable translation of Hindi text to Tamil text.

[2] https://tdil-dc.in/nplt/index.php?route=product/category&path=59

[3] https://www.microsoft.com/en-us/download/details.aspx?id=105292

[4] https://opennmt.net/

Table 2: Performance of transformer-based NMT model and effect of parameter tuning

Test Dataset	NMT Model	BLEU	BERTScore	SBSim
918 sentences	Baseline	11.27	83.03	80.02
	FF-4096	11.77	83.06	81.20
	Attn-16	11.27	82.69	79.98
	W2V-1048	9.69	81.84	78.59
2000 sentences	Baseline	16.46	84.55	83.01
	FF-4096	16.70	84.62	83.98
	Attn-16	17.27	84.50	83.05
	W2V-1048	14.46	83.50	81.86

The baseline transformer-based model parameters trained in the current work is given in Table 1. The model is trained using the Hindi-Tamil Samanantar dataset discussed in Section 2. Each of these parameters are crucial and tuning them affects the performance of the end translation model. In the current work, we tweak few of these parameters (as seen in Table 1) and observe their effect on the translation performance.

The first change made is the size of the feed-forward (ff) layer in the network. While the baseline model's ff-layer size is 2048, the tweaked model, FF-4096, is set to have an ff-layer size of 4098. In addition to this, the number of attention heads is increased to 16 (model Attn-16) as compared to the baseline model with 8 attention heads. In model W2V-1028, the word2vec size and the hidden layer vector size is set to 1028 as compared to 512 in the base model.All these changes in the parameters are expected to better represent the training data, extract complex relationships and information, and result in a better translation model. The performance of the baseline and the modified models are tabulated in Table 2.

From the Table it can be observed that, the performance of the model improves when the feed forward parameter is increased to 4096. However, in case of the user-defined test set (918 sentences) increasing the attention head is observed to degrade the translation performance. This maybe due to the fact that the user-defined dataset contains small length and conversational sentences which the model is not accustomed to. Similarly, increasing the vector sizes reduces the performance for both the test sets. This is likely due to the overfitting caused by the vector representations. An example to validate these observations is given below.

Input: यहाँ कौन सा अच्छा होटल है?

Google reference: இங்கே எது நல்ல ஹோட்டல்?

Baseline: என்ன நல்ல ஹோட்டல்?

FF-4096: இங்க என்ன நல்ல ஹோட்டல் இருக்கு?

Attn-16: என்ன நல்ல விடுதி அது?

W2v-1048: எந்த ஹோட்டல்?

From the above example it can be observed that, FF-4096 is the most appropriate translation for the input sentence without insertions or deletions of words as observed in the other system outputs.

Thus, it is optimum to increase the feed forward size of the baseline network while keeping the remaining parameters untouched. With these inferences from the model architecture the next set of experiments are carried out to determine the effect of finetuning on MT models.

4 Effect of finetuning with domain specific data

Artificial Intelligence (AI) Systems for Indic languages has been a much sought after initiative in recent times and AI4Bharat [5] provides several NLP and speech related applications for Indic languages. One such application is the IndicTrans model which is a transformer-based multilingual NMT model trained on the Samanantar dataset for 22 scheduled Indian languages (Dixit et al., 2023). In the current work, this pretrained model (also called m2m model) is used as the base model. As inferred from the previous experiments, the

[5] https://ai4bharat.iitm.ac.in/

Table 3: Effect of finetuning on the performance of NMT model

Test Dataset	Model	BLEU	BERTScore	SBSim
918 sentences	m2m	17.75	86.98	87.71
	m2m+TDIL	19.25	87.19	88.02
	m2m+TDIL+Microsoft	21.53	88.13	89.08
2000 sentences	m2m	23.09	87.10	86.93
	m2m+TDIL	24.14	88.60	87.13
	m2m+TDIL+Microsoft	26.84	90.08	89.52

m2m model is also observed to have a feed-forward size of 4096 and an attention head of 16.

As discussed in the previous section, the performance of the system maybe affected due to unavailability of short and conversational sentences in the training data. Thus, finetuning a parent model to suit the needs of the final application is expected to result in better performance. On this note, the pretrained model is finetuned for better performance using the TDIL and Microsoft dataset mentioned in Section 2.

The effect of the finetuning on the performance of the translation model is tabulated in Table 3. From the Table it is observed that, for both the test datasets, finetuning the m2m model with TDIL data improves scores by a minimum of 1% to a maximum of 2%. Further, finetuning this model with Microsoft data improves the performance by a minimum of 1% to a maximum of 2%. Thus, an overall improvement of 4% is observed upon finetuning the m2m model. In order to further validate these observations an example translation is provided below.

Input: आपकी और क्या विशेषता है?

Google reference: உங்களுக்கு வேறு என்ன சிறப்பு?

m2m: உங்களின் தனித்துவம் என்ன?

m2m+TDIL உங்களுக்கு வேறு என்ன சிறப்பு உள்ளது?

m2m+TDIL+microsoft உங்களுக்கு வேறு என்ன சிறப்பு இருக்கிறது?

The above example shows that the m2m output considers विशेषता as 'uniqueness' and translates it to தனித்துவம், while the m2m+TDIL and m2m+TDIL+microsoft consider it as 'speciality' and translate it to சிறப்பு. Further, the m2m+TDIL+microsoft makes use of இருக்கிறது instead of உள்ளது to end the sentence making it more conversational.

Thus, from the above experiments it can be inferred that the best approach to develop a domain-specific conversational machine translation system is to tweak certain hyperparamters of the model followed by finetuning a pretrained model with the domain data.

5 Conclusion

This paper aims to develop a domain-specific translation system for Hindi-to-Tamil translation. The paper focuses on analysing the effect of hyper-parameter tuning and finetuning a pretrained model with a domain-specific dataset. It is observed that increasing the feed-forward size, and attention heads in the transformer MT model improves the translation performance. In addition to this, finetuning the model with conversational and travel-domain specific dataset improves the conversational nature of the translated outputs. In conclusion, experimental results demonstrate that combining domain-specific finetuning with optimal parameter settings significantly enhances translation performance, making the system more efficient for industry applications. The future of this work can be directed towards including more conversational features or data to make the system suitable for common use.

Limitations

The limitations of the paper is predominantly associated to the availability of clean, and large dataset to train and model the translation systems. Though the amount of text available is large, the quality of the text is not appreciable and requires lot of manual intervention to clean it. Collecting data from users is also a challenge as it is essential to locate

the right set of people and to extract the right kind of text from them. Further, the computational requirements, such as GPU power, to experiment and investigate various network architectures and parameters is limited making the process of experimentation laborious and time-consuming.

In addition to the data and system requirements, translation errors such as negative translation, incomplete translation do exist. This can be improved by including more data for fine-tuning and incorporating language features such as POS tagging and morphological analysis. Also code-switching (i.e., input sentences containing both Hindi and English words) is to be addressed to ensure that English words are not translated. This can be done using language tags.

Ethics Statement

The authors confirm that the current work complies with the ACL ethics policy.

Acknowledgements

The authors would like to thank DST-SERB IMPRINT II C, for funding the project titled "Standalone Domain Specific Speech-to-Speech Translator for English, Hindi and Tamil Languages". Ref. No. IMC/2020/000003-G.

References

Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, et al. 2023. Indicmt eval: A dataset to meta-evaluate machine translation metrics for indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228.

K. Mrinalini and P. Vijayalakshmi. 2015. Hindi-english speech-to-speech translation system for travel expressions. In *2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC)*, pages 0250–0255.

K. Mrinalini, P. Vijayalakshmi, and T. Nagarajan. 2022. Sbsim: A sentence-bert similarity-based evaluation metric for indian language neural machine translation systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1396–1406.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.