Attentive fine-tuning of Transformers for Translation of low-resourced languages
@LoResMT 2021

Karthik Puranik1 Adeep Hande1 Ruba Priyadharshini2 Thenmozi Durairaj3
Anbukkarasi Sampath4 Kingston Pal Thamburaj5 Bharathi Raja Chakravarthi6
1Department of Computer Science, Indian Institute of Information Technology Tiruchirappalli
2ULTRA Arts and Science College, Madurai, Tamil Nadu, India
3Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu, India
4Kongu Engineering College, Erode, Tamil Nadu, India
5Sultan Idris Education University, Tanjong Malim, Perak, Malaysia
6Insight SFI Research Centre for Data Analytics, National University of Ireland Galway

Abstract

This paper reports the Machine Translation (MT) systems submitted by the IIITT team for the English→Marathi and English↔Irish language pairs LoResMT 2021 shared task. The task focuses on getting exceptional translations for rather low-resourced languages like Irish and Marathi. We fine-tune IndicTrans, a pretrained multilingual NMT model for English→Marathi, using external parallel corpus as input for additional training. We have used a pretrained Helsinki-NLP Opus MT English↔Irish model for the latter language pair. Our approaches yield relatively promising results on the BLEU metrics. Under the team name IIITT, our systems ranked 1, 1, and 2 in English→Marathi, Irish→English, and English→Irish respectively. The codes for our systems are published1.

1 Introduction

Today, a large number of text and written materials are present in English. However, with roughly around 6,500 languages in the world2 (Chakravarthi, 2020; Hande et al., 2021a; Sarveswaran et al., 2021), every native monoglot should not be deprived of this knowledge and information. The manual translation is a tedious job involving much time and human resources, giving rise to Machine Translation (MT). Machine Translation involves the automated translation of text from one language to another by using various algorithms and resources to produce quality translation predictions (Pathak and Pakray, 2018; Krishnamurthy, 2015, 2019). Neural Machine Translation (NMT) brought about a great improvement in the field of MT by overcoming flaws of rule-based and statistical machine translation (SMT) (Revanuru et al., 2017; Achchuthan and Sarveswaran, 2015; Parameswari et al., 2012; Thenmozhi et al., 2018; Kumar et al., 2020b). NMT incorporates the training of neural networks on parallel corpora to predict the likeness of a sequence of words. sequence-to-sequence neural models (seq2seq)

1https://github.com/karthikpuranik11/LoResMT
2https://blog.busu.com/most-spoken-languages-in-the-world/
(Sutskever et al., 2014; Kalchbrenner and Blunsom, 2013) are the widely adopted as the standard approach by both industrial and research communities (Jadhav, 2020a; Bojar et al., 2016; Cheng et al., 2016).

Even though NMT performs exceptionally well for all the languages, it requires a tremendous amount of parallel corpus to produce meaningful and successful translations (Kumar et al., 2020a). With little research on low resourced languages, finding a quality parallel corpus to train the models can be arduous. The two low-resourced languages worked on in this paper are Marathi (mr) and Irish (ga). With about 120 million Marathi speakers in Maharashtra and other states of India, Marathi is recognized as one of the 22 scheduled languages of India\(^3\). The structural dissimilarity which occurs while translating from English (Subject-Verb-Object) to Marathi (Subject-Object-Verb) or vice versa adds up to issues faced while translation (Garje, 2014). The Irish language was recognized as the first official language of Ireland and also by the EU (Dowling et al., 2020a; Scannell, 2007). Belonging to the Goidelic language family and the Celtic family (Scannell, 2007; Lynn et al., 2015), Irish is also claimed as one of the low resourced languages due to its limited resources by the META-NET report (Dhonnchadha et al., 2012; Scannell, 2006).

Our paper represents the work conducted for the LoResMT @ MT Summit 2021\(^4\) shared task to build MT systems for the low-resourced Marathi and Irish languages on COVID-19 related parallel corpus. We implement Transformer-based (Vaswani et al., 2017) NMT models to procure BLEU scores (Papineni et al., 2002) of 24.2, 25.8, and 34.6 in English $\rightarrow$ Marathi, Irish $\rightarrow$ English, and English $\rightarrow$ Irish respectively.

2 Related works

Neural Machine Translation has been exhaustively studied over the years (Kalchbrenner and Blunsom, 2013), with several intuitive approaches involving collective learning to align and translate (Bahdanau et al., 2016), and a language-independent attention bridge for multilingual translational systems (Vázquez et al., 2019). There have been several approaches to NMT, with zero-shot translational systems, between language pairs that have not seen the parallel training data during training (Johnson et al., 2017). The introduction of artificial tokens has reduced the architectural changes in the decoder (Ha et al., 2016). There have been some explorations towards neural machine translation in low resource languages, with the development of a multi-source translational system that targets the English string for any source language (Zoph and Knight, 2016).

There has been subsequent research undertaken by researchers for machine translation in low-resource Indian languages. Chakravarthi et al. surveyed orthographic information in machine translation, examining the orthography’s influence on machine translation and extended it to under-resourced Dravidian languages (Chakravarthi et al., 2019a). Another approach of leveraging the information contained in rule-based machine translation systems to improve machine translation of low-resourced languages was employed (Torregrosa et al., 2019). Several approaches involving the improvement of WordNet for low-resourced languages have been explored (Chakravarthi et al., 2018, 2019b). Chakravarthi et al. constructed MMDravi, a multilingual multimodal machine translation dataset for low-resourced Dravidian languages, extending it from the Flickr30K dataset, and generating translations for the captions using phonetic transcriptions (U Hegde et al., 2021).

There have been relatively fewer approaches experimented with and benchmarked when it comes to translating from Marathi to English and vice versa. (Aharoni et al., 2019) tried to build multilingual NMT systems, comprising 103 distinct languages and 204 translational directions.

\(^3\)https://en.wikipedia.org/wiki/Marathi_language
\(^4\)https://sites.google.com/view/loresmt/
simultaneously. (Jadhav, 2020b; Puranik et al., 2021) developed a machine translation system for Marathi to English using transformers on a parallel corpus. Other works on improving machine translation include (Adi Narayana Reddy et al., 2021) proposing an English-Marathi NMT using local attention. The same can be stated to Irish, as it is a poorly resourced language, as the quality of the MT outputs have struggled to achieve the same level as well-supported languages (Dowling et al., 2016; Rehm and Uszkoreit, 2012). In recent years, several researchers tried to overcome the resource barrier by creating artificial parallel data through back-translation (Poncelas et al., 2018), exploiting out-of-domain data (Imankulova et al., 2019), and leveraging other better-resourced languages as a pivot (Dowling et al., 2020b; Wu and Wang, 2007).

3 Dataset

We use the dataset provided by the organizers of LoResMT @ MT Summit 2021. The datasets can be found here\(^5\). It is a parallel corpus for English and the low resourced language, i.e., Irish and Marathi, mostly containing text related to COVID-19 (Ojha et al., 2021).

<table>
<thead>
<tr>
<th>Language pair</th>
<th>English⇔Irish</th>
<th>English⇔Marathi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>8,112</td>
<td>20,933</td>
</tr>
<tr>
<td>Dev</td>
<td>502</td>
<td>500</td>
</tr>
<tr>
<td>Test</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Total</td>
<td>9,614</td>
<td>22,433</td>
</tr>
</tbody>
</table>

Table 1: Number of sentences distribution

We have used bible-uedin\(^6\) (Christodoulopoulos and Steedman, 2015) an external dataset for Marathi. It is a multilingual parallel corpus dataset containing translations of the Bible in 102 languages (Christodoulopoulos and Steedman, 2014) and shows the possibility of using the Bible for research and machine translation. English-Marathi corpus contains 60,495 sentences. CVIT PIB\(^7\) (Philip et al., 2020) has also been used for the purpose of this research. It contains 1,14,220 parallel corpora for English-Marathi.

4 Methodology

4.1 IndicTrans

Fairseq PyTorch\(^8\) (Ott et al., 2019) is an open-source machine learning library supported as a sequence modeling toolkit. Custom models can be trained for various tasks, including summarization, language, translation, and other generation tasks. Training on fairseq enables competent batching, mixed-precision training, multi-GPU and multi-machine training. IndicTrans (Ramesh et al., 2021), a Transformer-4x multilingual NMT model by AI4Bharat, is trained on the Samanantar dataset. The architecture of our approach is displayed in Fig.1. Samanantar\(^9\) is the most extensive collection of parallel corpora for Indic languages available for public use. It includes 46.9 million sentence pairs between English and 11 Indian languages. IndicTrans is claimed to successfully outperform the existing best performing models on a wide variety of benchmarks. Even commercial translation systems and existing publicly available systems were surpassed for the majority of the languages. IndicTrans is based on fairseq, and it was

\(^5\)https://github.com/loresmt/loresmt-2021
\(^6\)https://opus.nlpl.eu/JRC-Acquis.php
\(^7\)http://preon.iiit.ac.in/~jerin/bhasha/
\(^8\)https://github.com/pytorch/fairseq
\(^9\)https://indicnlp.ai4bharat.org/samanantar/
Figure 1: Our approach for the English→Marathi language pair.

The model was fine-tuned on the Marathi training dataset provided by the organizers and the external datasets. The model was fine-tuned with the cross-entropy criterion to compute the loss function, Adam optimizer (Zhang, 2018), dropout of 0.2, fp16 (Micikevicius et al., 2017), maximum tokens of 256 for better learning, and a learning rate if 3e-5 in GPU. The process was conducted for a maximum of 3 epochs.

4.2 Helsinki-NLP Opus-MT

OPUS-MT (Tiedemann and Thottingal, 2020) supports both bilingual and multilingual models. It is a project that focuses on the development of free resources and tools for machine translation. The current status is a repository of over 1,000 pretrained neural MT models. We fine-tune a transformer-align model that was fine-tuned for the Tatoeba-Challenge\textsuperscript{10}. Helsinki-NLP/opus-mt-en-ga model from the HuggingFace Transformers (Wolf et al., 2020) for English→Irish and Helsinki-NLP/opus-mt-ga-en for Irish→English were used.

<table>
<thead>
<tr>
<th>Language pair</th>
<th>Method</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>English→Marathi</td>
<td>IndicTrans baseline</td>
<td>14.0</td>
</tr>
<tr>
<td>English→Marathi</td>
<td>IndicTrans TRA</td>
<td>17.8</td>
</tr>
<tr>
<td>English→Marathi</td>
<td>IndicTrans CVIT-PIB</td>
<td>23.4</td>
</tr>
<tr>
<td>English→Marathi</td>
<td>IndicTrans bible-uedin</td>
<td>27.7</td>
</tr>
<tr>
<td>English→Irish</td>
<td>Opus MT</td>
<td>30.4</td>
</tr>
<tr>
<td>English→Irish</td>
<td>M2M100</td>
<td>25.6</td>
</tr>
<tr>
<td>Irish→English</td>
<td>Opus MT</td>
<td>37.2</td>
</tr>
<tr>
<td>Irish→English</td>
<td>M2M100</td>
<td>30.4</td>
</tr>
</tbody>
</table>

Table 2: BLEU scores obtained for the various models for the development set

5 Results and Analysis

For Marathi, it is distinctly visible that our system model, i.e., IndicTrans fine-tuned on the training data provided by the organizers or TRA and the bible-uedin dataset, gave the best BLEU

\textsuperscript{10}https://github.com/Helsinki-NLP/Tatoeba-Challenge
scores. It was surprising how the model fine-tuned on a parallel corpus of 60,495 sentences of bible-uedin surpassed the model fine-tuned on 1,14,220 sentences from the CVIT PIB dataset. The possible explanation is the higher correlation between the sentences of the bible-uedin dataset with the test dataset than the CVIT PIB dataset. Another reason could be the presence of excessive noise in the CVIT PIB dataset. The other reason for this could be noise and a lower quality of translations in the CVIT PIB dataset compared to bible-uedin.

To infer the same, 1000 random pairs of sentences were picked from the datasets, and the average LaBSE or language-agnostic BERT Sentence Embedding (Feng et al., 2020) scores were found out. LaBSE gives a score between 0 and 1, depending on the quality of the translation. It was seen that the average was 0.768 for the bible-uedin dataset, while it was 0.58 for the CVIT PIB dataset. This might have been one of the reasons for the better BLEU scores. The model also showed constant overfitting after the second and third epoch, as the BLEU scores reduced considerably as they reached the 6th epoch. The BLEU scores decreased by a difference of 6. The validation loss starts increasing after the third epoch, thus, showing the overfitting occurring in training. So, the model was fine-tuned for three epochs while maintaining a low learning rate of around 3e-5 to get a BLEU score of 24.2.

<table>
<thead>
<tr>
<th>Language pair</th>
<th>BLEU</th>
<th>CHRF</th>
<th>TER</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>English→Marathi</td>
<td>24.2</td>
<td>0.59</td>
<td>0.597</td>
<td>1</td>
</tr>
<tr>
<td>Irish→English</td>
<td>34.6</td>
<td>0.61</td>
<td>0.711</td>
<td>1</td>
</tr>
<tr>
<td>English→Irish</td>
<td>25.8</td>
<td>0.53</td>
<td>0.629</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3: Result and ranks obtained for the test dataset (Popović, 2015; Snover et al., 2006)

Training a model to predict for a low-resourced language was highly challenging due to the absence of prominent pretrained models (Kalyan et al., 2021; Yasaswini et al., 2021; Hande...
et al., 2021b). However, as an experiment, two models from HuggingFace Transformers\textsuperscript{11}, M2M100 (Fan et al., 2020) and Opus-MT from Helsinki NLP (Tiedemann, 2020) were compared. For the dev data, Opus MT produced a BLEU score of 30.4 while M2M100 gave 25.62 for translations from English to Irish and 37.2 and 30.37 respectively for Irish to English translations. Probably, the individual models pretrained on numerous datasets gave Opus MT an edge over M2M100. This led us to submit the Opus MT model for the LoResMT Shared task 2021. The model gave exceptional BLEU scores of 25.8 for English to Irish, which ranked second in the shared task, while 34.6 for Irish to English stood first.

6 Conclusion

It is arduous and unyielding to get accurate translations for low-resourced languages due to limited datasets and pretrained models. However, our paper puts forward a few methods to better the already existing accuracies. Ranked 1, 1, and 2 in English→Marathi, Irish→English, and English→Irish respectively in the LoResMT 2021 shared task, IndicTrans fine-tuned on the bible-uedin, and the dataset provided by the organizers manages to surpass the other models due to its high correlation with the test set and minimal noise for the Marathi language. The Irish language task was dominated by the Opus MT model by Helsinki-NLP, outperforming other Transformer models, M2M100.

References


\textsuperscript{11}https://huggingface.co/transformers/


Philip, J., Siripragada, S., Namboodiri, V. P., and Jawahar, C. V. (2020). Revisiting low resource status of indian languages in machine translation. 8th ACM IKDD CODS and 26th COMAD.


