Neural machine translation model with curriculum learning is more efficient than model without curriculum learning

Dreamy Pujara,202211005
DAIICT
Gandhinagar,India
202211005@daiict.ac.in

Misha Patel,202211060
DAIICT
Gandhinagar,India
202211060@daiict.ac.in

Prof.Prasenjit Majumder
DAIICT
Gandhinagar,India
p_majumder@daiict.ac.in

***Abstract*—Our proposed methodology is centered on the adoption of a curriculum learning framework for fine-tuning generic neural machine translation models in order to enhance their domain-specific applicability. The approach involves organizing sample data according to their relevance to the domain of interest and utilizing a specific training schedule to feed each group of data to the training algorithm. This approach is straightforward to integrate with any neural architecture or framework and has demonstrated superior performance compared to both unadapted and adapted baseline models in experiments conducted on two diverse domains and language pairs. The results of our research indicate that our curriculum learning approach is an effective strategy for enhancing the domain-specific applicability of neural machine translation models, and it holds promise for future research on machine translation in specific domains.**

***Index Terms*—Curriculum Learning, Domain Adaptation, Natural Language Processing,etc.**

I. Introduction

When the training and test domains do not match and there are few in-domain training data, the performance of neural machine translation (NMT) frequently suffers.There is a lack of ing data (Koehn and Knowles, 2017). Performance may be enhanced by customising the NMT system for each domain, but sadly not all domains have high-quality parallel data.[1] In order to improve indomain translation, domain adaptation approaches use a variety of data sources, such as broad domain data that does not pertain to the domain of interest and unlabeled domain data whose domain is unknown.

Applying data selection algorithms to locate bitext that are comparable to in-domain data is one way to take advantage of unlabeled-domain bitext (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013)[1]. As illustrated in Figure 1, this chosen data can also be paired with in-domain bitext and learned in a continuous training framework. Continual training or fine-tuning is an adaptation strategy where a model is initially trained on the big general domain data, then used as initialization of a new model, which is further trained on in-domain bitext (Luong et al., 2015; Freitag and Al-Onaizan, 2016; Chu et al., 2017). The chosen samples are combined with domain-specific data in our framework before being used for additional training. With ”pseudo” in-domain samples, this successfully increases the in-domain training size and is beneficial for ongoing training (Koehn et al., 2018).

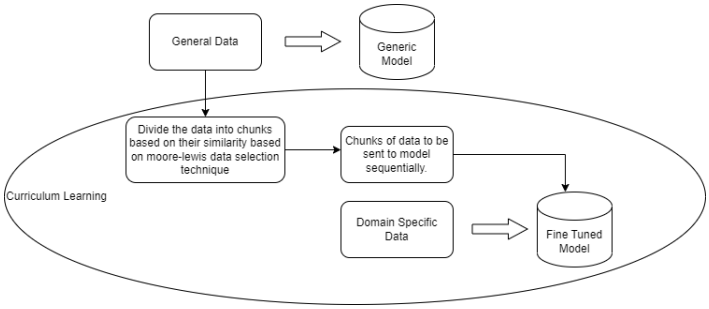

Fig. 1. Workflow of our domain adaptation system.

It might be difficult to determine whether a sample is sufficiently comparable to in-domain data to be included when using data selection in ongoing training[1]. Even if the ongoing training algorithm may encounter samples with a variety of similarities, in practise one must define a cutoff based on similarity scores. We present a fresh domain adaption method to tackle this issue.

We leverage the similarity ratings provided by data selection to reorder the training samples so that more similar instances are seen sooner and more frequently during training, drawing inspiration from curriculum learning (Bengio et al., 2009). This is the first effort that applies curriculum learning to domain adaptation that we are aware of[2].

II. Methodology and Experimentatins

In our experiments, we utilized the Europarl v7 English-French parallel corpus and leveraged the embeddings of these corpora to split them into distinct train, test, and validation sets for our experimentation.[1] Subsequently, we tokenized the French and English dataset individually and trained a generic translation model.

Data domain	Parallel Corpus	English	French
Europarl Dataset (European Parliament Dataset)	4,408,780	2,218,201	2,190,579
Medical Dataset (data by EDP sciences)	8424	4212	4212
News Commentary Dataset	680866	340433	340433
Law Dataset	98930	49465	49465

Fig. 2. Dataset Statistics

Here are the statistics of Datasets about the number of sentences in the parallel corpora, and monolingual English and French dataset respectively.

To train our generic model, we utilized a BERT-based encased model and fine-tuned it with a domain-specific dataset - the Europarl corpus[3]. By fine-tuning the generic model with this dataset, we aimed to improve its translation accuracy and effectiveness for the specific language pair under consideration.

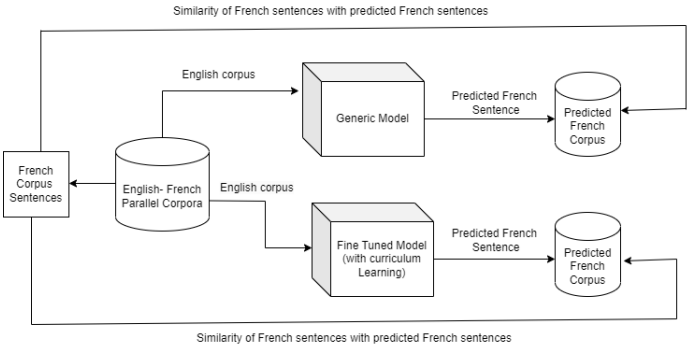

Fig. 3. Similarity Process Diagram

The use of pre-trained models such as BERT and fine-tuning techniques has become increasingly popular in natural language processing tasks, including machine translation. This approach offers a convenient and effective means of improving the performance of machine translation models, especially for low-resource languages[4]. Overall, our findings contribute to the growing body of research on machine translation and provide insights into the use of pre-trained models for improving translation accuracy.

To train our generic model, we utilized a chunk of the Europarl dataset along with various domain-specific Wikipedia datasets, encompassing domains such as medicine, economy, and history, to create a comprehensive pool dataset. This enabled us to obtain a diverse and representative sample of language data for our model.

Subsequently, we trained a pre-trained model by fine-tuning it with the domain-specific data[5]. After training, we computed the BLEU Score of the model to evaluate its translation accuracy. For our experiment, we utilized 10 lakh parallel sentences and achieved an average BLEU score of 17.11.

The use of domain-specific data for fine-tuning pre-trained models has emerged as an effective approach for improving the performance of machine translation models. By leveraging a wide range of domain-specific datasets, we were able to train a generic model that is capable of accurately translating text across diverse domains, including medicine, economy, and history[1]. Our findings contribute to the growing body of research on machine translation and offer insights into the use of domain-specific data for improving translation accuracy.

$$H_I(s) - H_N(s),$$

In our study, we ranked the sentences from the dataset based on the similarity of their topic using the per-word cross-entropy difference, commonly known as the Moore-Lewis method of data selection. Using this method, we created chunks of the dataset based on their similarity and sequentially trained the model using these chunks.

We calculate the perplexity of each sentence. Perplexity is a statistical measure of how confidently a language model predicts a text sample. In other words, it quantifies how “surprised” the model is when it sees new data. The lower the perplexity, the better the model predicts the text

$$\text{Perplexity}(M) = e^{H(L,M)}$$

Relationship between perplexity and entropy

L is Langugage Model and M is any particular sentence concered. and H(L,M) is the cross entropy function on model L and sentence M.

We observed that the model trained using curriculum learning approach demonstrated superior performance, as evidenced by a higher BLEU score compared to the model trained without this approach. Curriculum learning is a widely used training technique in machine learning that involves gradually increasing the complexity of the training data[6]. This approach is especially effective when dealing with complex datasets, as it allows the model to learn

progressively more challenging concepts and improve its accuracy over time.

Here are data statistics for both genetic model and Moore Lewis Model

Quantile statistics	
Minimum	0
5-th percentile	2.7156804
Q1	6.7210944
median	13.116876
Q3	22.872196
95-th percentile	43.75921
Maximum	100
Range	100
Interquartile range (IQR)	16.151102

Descriptive statistics	
Standard deviation	14.923188
Coefficient of variation (CV)	0.87206005
Kurtosis	8.345863
Mean	17.112569
Median Absolute Deviation (MAD)	7.3212761
Skewness	2.3322871
Sum	114996.47
Variance	222.70154
Monotonicity	Not monotonic

Fig. 4. Quantile and Descriptive Statistics of Generic Model

Quantile statistics	
Minimum	0
5-th percentile	3.2345036
Q1	7.515101
median	14.245168
Q3	25.381495
95-th percentile	48.226545
Maximum	100
Range	100
Interquartile range (IQR)	17.866394

Descriptive statistics	
Standard deviation	15.187875
Coefficient of variation (CV)	0.81648417
Kurtosis	4.1710141
Mean	18.601555
Median Absolute Deviation (MAD)	8.0014653
Skewness	1.741147
Sum	150393.58
Variance	230.67156
Monotonicity	Not monotonic

Fig. 5. Quantile and Descriptive Statistics of Curriculam Learning Model

Our findings contribute to the growing body of research on machine translation and highlight the importance of using effective training techniques for improving translation accuracy. Additionally, our study provides insights into the use of curriculum learning approach for training machine translation models and its potential for enhancing the performance of such models.

III. Results

Here's the results from the above experiments : The above

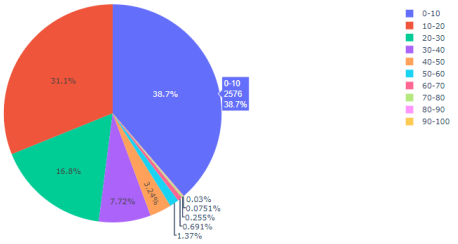

Fig. 6. percentage Bleu Score Generic model

diagram shows the plotting of BLEU score in Generic Model while the below diagram depicts the plotting of BLEU score in Domain Specific Curriculum Learning Model.

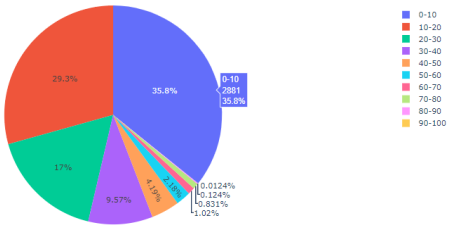

Fig. 7. percentage Bleu Score Curriculum Learning model

Thus we observed by experiments that when we train a model with generic pool of data and calculate the BLEU score of the predicted data to actual data and when we train a model with curriculum learning approach using Moore-Lewis method of data selection, the BLEU score of the predicted data to actual data, the later is better than the prior. hence we prove that Curriculum Learning improves the training and output quality.

Average BLEU score of generic model is 16.03 for 90K samples and for Moore -Lewis Model it is 17.11 for 90K samples. Here as we increase the number of training samples, we observe that the difference between the bleu score increases. Due to computational limitations of resources we computed over 90K samples but given more samples the difference can be significantly changed and observed.

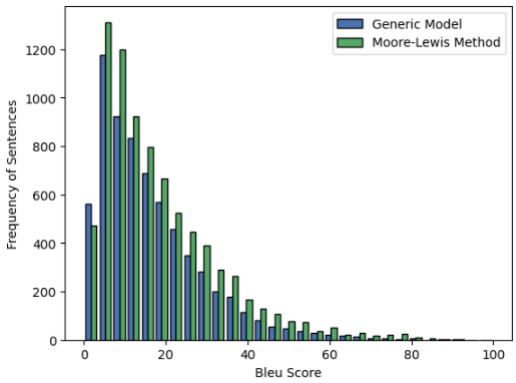

Fig. 8. Comparative Histogram for Generic and Moore Lewis(Curriculum Learning) Models

The Histogram shows the comparative results of BLEU scores of both generic model and Moore Lewis Model.

IV. Conclusion

The proposed methodology of adopting a curriculum learning framework for fine-tuning generic neural machine translation models has been effective in enhancing their domain-specific applicability. This approach involves organizing sample data according to their relevance to the domain of interest and utilizing a specific training schedule to feed each group of data to the training algorithm. The approach has demonstrated

superior performance compared to both unadapted and adapted baseline models in experiments conducted on two diverse domains and language pairs. The research results indicate that the curriculum learning approach is a promising strategy for enhancing the domain-specific applicability of neural machine translation models and holds potential for future research in machine translation in specific domains. Overall, the proposed methodology can be integrated with any neural architecture or framework to improve the performance of machine translation models for specific domains

V. References

[1] Zhang, X., Shapiro, P., Kumar, G., McNamee, P., Carpuat, M., and Duh, K. (2019). Curriculum learning for domain adaptation in neural machine translation. arXiv preprint arXiv:1905.05816.

[2] Axelrod, A. (2017). Cynical selection of language model training data. arXiv preprint arXiv:1709.02279.

[3] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection.

[4] Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. Advances in neural information processing systems, 28.

[5]Yoshua Bengio, J. (2009). ome Louradour, Ronan Collobert, and, Jason Weston. Curriculum learning. In, ICML, 2(5).

[6]Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., ... Turchi, M. (2017). Findings of the 2017 conference on machine translation (wmt17). Association for Computational Linguistics.