A Deep Transfer Learning Method for Cross-Lingual Natural Language Inference

Dibyanayan Bandyopadhyay[1], Arkadipta De[2], Baban Gain[1], Tanik Saikh[1] and Asif Ekbal[1]*

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna

[2] Department of Artificial Intelligence, Indian Institute of Technology Hyderabad

*Corresponding author: {dibyanayan_2111cs02}@iitp.ac.in

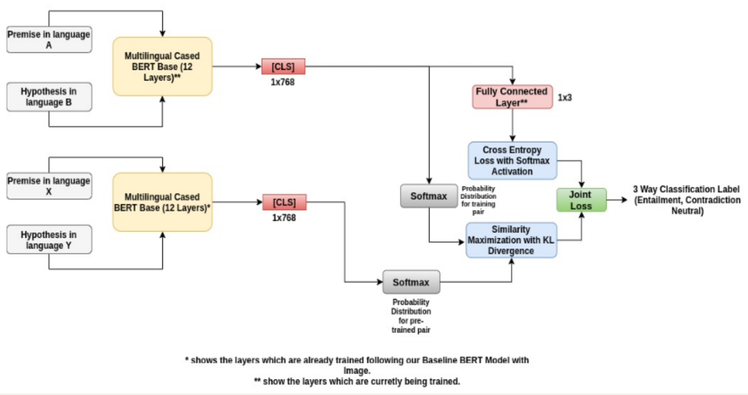
Introduction

Natural Language Inference (NLI) is an important task in the field of NLP. Particularly when it involves two separate languages as premise and hypothesis (called Cross-Lingual NLI). We aim to solve Cross-lingual NLI by proposing a novel loss formulation on top of existing architectures. It uses embedding learned by the hidden layers of a frozen cross-lingual model (e.g. trained on English-Hindi) to generalise it over a secondary model (e.g., trained on Bulgarian-French). We show that this method generalises all the combinations of four language pairs, namely French, German, Bulgarian and Turkish, on top of the XNLI dataset. We further show the hidden state dynamics to explicitly illustrate the learning behaviour of our model. We also compare our method with the standard knowledge distillation technique due to the similarity between the two methods. Experimental results verify our hypothesis that using our proposed method can be beneficial for Cross-lingual NLI rather than using standard knowledge distillation alone. We also compare our model with the state-of-the-art (XLM-R model) and it is shown that our model, despite a much lower parameter count than the SOTA, performs at per or even better.

Methodology: Using learned embedding to learn Cross-lingual NLI.

Cross-lingual NLI focuses on detecting the relationship between two pairs of text in different languages. To exploit a trained model, the representation space shared by various language embeddings should have to be similar. Therefore we choose multilingual BERT as our text encoder. Let's say an instance of multilingual BERT is trained on the English-Hindi language pair. We can subsequently use this trained model (henceforth called the teacher model T) to train another model S (henceforth called the student model) on other language pairs different from English-Hindi (e.g. Bengali-Assamese).

In our proposed approach, we materialise this by incorporating a KL divergence-based loss between the hidden representation of the teacher and the student models. Besides optimising the cross-entropy loss, model S tries to keep its hidden representation probabilistically as equal as possible to the hidden representation of T. This simple strategy is schematically described in the following Picture (a). Picture (b). illustrates the algorithm behind the proposed training methodology



(a). Overall architecture for our model

L = CrossEnt(a,b) + l*KL(h1,h2) is the loss that we use to train model S.

- 'a' and 'b' are output label and gold standards respectively.
- 'h1' and 'h2' are hidden representations of the teacher and student models respectively. They are obtained by using softmax distribution
- KL(h1,h2) is the KL divergence.
- 'l' is a hyper-parameter.

Experimental setup

- 'Baseline' and 'Proposed Model' both uses the same set of hyper-parameters.
- Batch Size = 28
- Learning Rate = 2e-5
- Maximum sequence length = 128
- Max epoch 5 with early stopping
- Random seed used is 42

Results

Model	Premise	Hypothesis	Accuracy (%)
Baseline System	English	Hindi	68.21
	Hindi	English	71.08
BERT-KLD	English	Hindi	70.36
	Hindi	English	72.3

Table 1: Results obtained on the baseline and the improvement on the baseline on different input modalities. *For all the experiments, λ=0.2*

Model	Language Pairs					
Baseline	*French-German*	*French-Turkish*	*French-Bulgarian*	*German-Turkish*	*German-Bulgarian*	*Turkish-Bulgarian*
	57.6	50	50.6	51.6	54.8	48

Table 2: Baseline System Performance across language pairs

Language Pairs	Transferring					
Transferred	*French-German*	*French-Turkish*	*French-Bulgarian*	*German-Turkish*	*German-Bulgarian*	*Turkish-Bulgarian*
French-German	-	50.4	57	50	57	51
French-Turkish	59	-	57	51	58	54.2
French-Bulgarian	59.2	52	-	51	57.6	53.8
German-Turkish	58.6	49.8	58.4	-	55.4	58.2
German-Bulgarian	57.8	51.8	56.8	49.4	-	54
Turkish-Bulgarian	58.8	51	57.2	49.6	57.2	-

Table 3: Performance of BERT-KLD for different transferring and transferred language pairs. We light the best score per column, which indicates which transferring language pairs are affecting the most the performance of the transferred language pairs.

Comparison with the SOTA

Models	Language Pairs					
	French-German	*French-Turkish*	*French-Bulgarian*	*German-Turkish*	*German-Bulgarian*	*Turkish-Bulgarian*
XLM-R	58.6	53.2	58.8	52.8	64.6	57.8
BERT-KLD	59.2	52	58.4	51	58	58.2

Table 4: Comparison of performance for our model (BERT-KLD) vs the state-of-the-art model (XLM-R)

Comparison with Knowledge Distillation

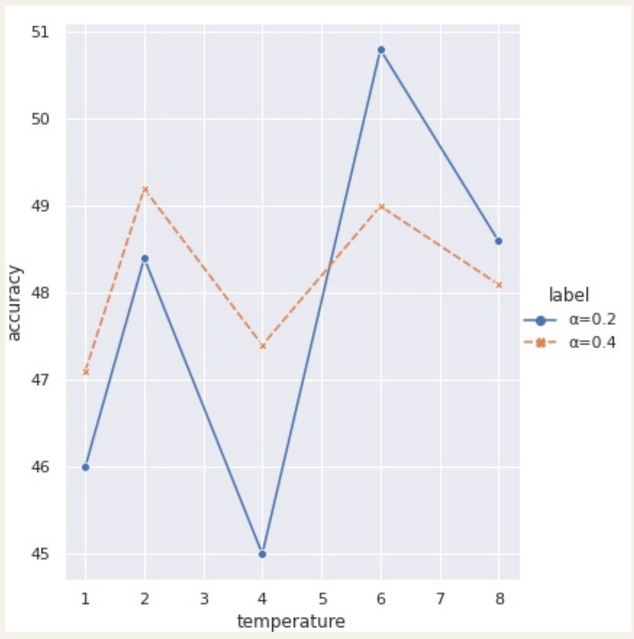
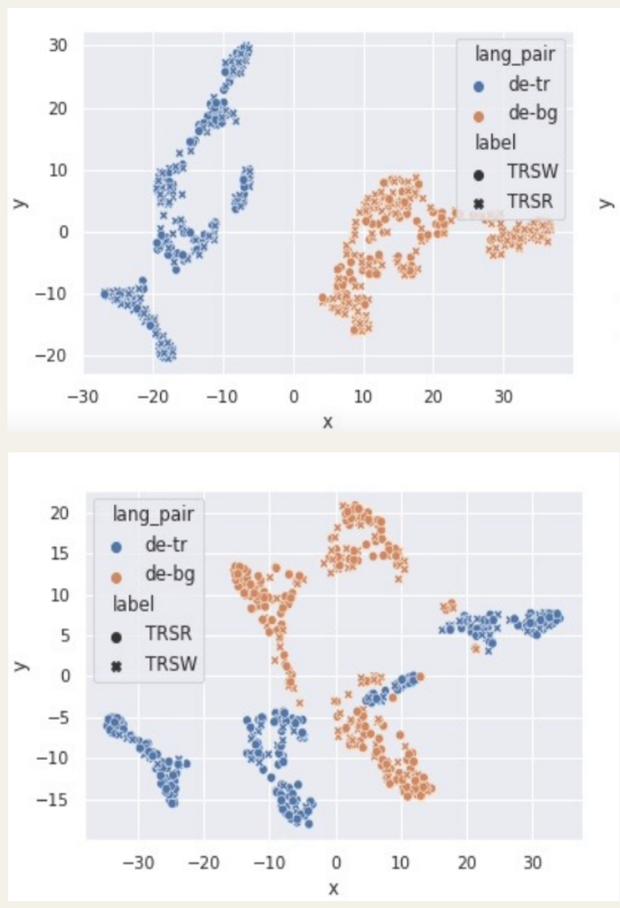


Figure 2: Performance of *BERT-KLD* employing classical KD framework. Transferring and transferred language pairs are *fr-bg* and *fr-tr* respectively. The peak accuracy obtained is 50.8%. α denotes the hyper-parameter associated with the extra loss term used to train the Student model. It is seen varying temperature parameters in softmax distribution can result in widely increased/decreased performance, with higher values of temperature usually giving better performance.

Dataset

- We have used EH-XNLI dataset [1] to evaluate the baseline and our proposed model in English-Hindi and Hindi-English setting.
- Further, we have used the XNLI dataset [2] to evaluate the baseline and BERT-KLD (our proposed model).
- For both datasets, our model performs much better in comparison to the baseline.
- For comparing against the SOTA, we choose XNLI set as our evaluation dataset.

Hidden State Analysis



Up: t-SNE projection of the hidden state vector of a trained baseline model before transfer on *de-tr* and *de-bg* language pairs.

Low: t-SNE projection of hidden state vector of both Teacher(P) and Student(Q) model after transfer learning. Before transfer took place (in the *left* figure), both of the baseline models have hidden states completely separate when projected on a 2-D plane. The upper figure demonstrates how the hidden state of both of those baseline models is close to each other when transferring knowledge from one to another. Teacher and Student models are trained on *de-tr* and *de-bg* respectively.

Conclusion and Future Works

We propose a novel transfer-learning algorithm to perform CLTE. We show the robustness of our method on by performing experiments on four European languages. We achieve state-of-the-art results in some language pairs in addition to having consistent improvements over baseline for all language pairs. We use the BERT model for classification by combining Cross-entropy as well as KL-divergence for similarity maximisation. We compare the results with existing CLTE systems. In future, we would like to extend our work toward Multimodal CLTE. We finally note that this training method is model agnostic and can be used as a plug and play replacement for classical knowledge distillation frameworks in downstream NLP tasks. In future, we want to explore in this direction to apply this methodology in other NLP tasks.

References

[1] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. Association for Computational Linguistics.

[2] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[3] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network.

[4] Saikh, T., De, A., Bandyopadhyay, D., Gain, B., and Ekbal, A. (2020). A neural framework for English- Hindi cross-lingual natural language inference. In Haiqin Yang, et al., editors, *Neural Information Processing - 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23-27, 2020, Proceedings, Part I*, volume 12532 of *Lecture Notes in Computer Science*, pages 655–667. Springer.