Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3084–3092
Marseille, 20-25 June 2022

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

Dibyanayan Bandyopadhyay[1] **Arkadipta De**[2*] **Baban Gain**[1]
Tanik Saikh[1] **Asif Ekbal**[1]
[1]Department of Computer Science and Engineering,
[1]Indian Institute of Technology Patna, India
[1]Department of Artificial Intelligence,
[2]Indian Institute of Technology Hyderabad, India
[1]{dibyanayan_2111cs02, baban_2111cs21, 1821cs08, asif}@iitp.ac.in
[2]{ai20mtech14002}@iith.ac.in

Abstract

Natural Language Inference (NLI), also known as Recognizing Textual Entailment (RTE), has been one of the central tasks in Artificial Intelligence (AI) and Natural Language Processing (NLP). RTE between the two pieces of texts is a crucial problem, and it adds further challenges when involving two different languages, i.e., in the cross-lingual scenario. This paper proposes an effective transfer learning approach for cross-lingual NLI. We perform experiments on English-Hindi language pairs in the cross-lingual setting to find out that our novel loss formulation could enhance the performance of the baseline model by up to 2%. To assess the effectiveness of our method further, we perform additional experiments on every possible language pair using four European languages, namely French, German, Bulgarian, and Turkish, on top of XNLI dataset. Evaluation results yield up to 10% performance improvement over the respective baseline models, in some cases surpassing the state-of-the-art (SOTA). It is also to be noted that our proposed model has 110M parameters which is much lesser than the SOTA model having 220M parameters. Finally, we argue that our transfer learning-based loss objective is model agnostic and thus can be used with other deep learning-based architectures for cross-lingual NLI.

Keywords: Natural Language Inference, Textual Entailment, Cross-lingual

1. Introduction

Textual Entailment (TE) is one of the fundamental problems of Natural Language Understanding (NLU). Understanding entailment and contradiction is fundamental to understanding natural language, and inference about entailment and contradiction is vital for developing semantic representations of natural languages. For two sentences, premise (P) and hypothesis (H), the Natural Language Inference (NLI) task is to understand the relationship between these two sentences (*viz.* Entailment, Contradiction, Unknown/Neutral). It is an elusive and frontier problem in Artificial Intelligence (AI). More specifically, H entails P if it strictly follows the statements:

- Hypothesis (H) is a logical consequence of Premise (P).
- Hypothesis (H) is true in every circumstance in which the Premise (P) is true.

If both P and H are in different languages, the problem is known as Cross-Lingual Textual Entailment (CLTE). Over the years, Natural Language Inference (NLI) has been addressed using a large variety of techniques, including those based on symbolic logic, knowledge bases, machine learning algorithms, and neural networks. While the field of Textual Entailment is quite popular and developed, limited research has been done in the field of CLTE. The main challenge of CLTE is that the fragments of texts are in different languages having different semantic and syntactic structures and grammar. NLI or TE aims to detect logical consequences within a given pair of text fragments, and to do this efficiently, we need to incorporate world knowledge and facts. In this paper, we focus on the challenging task of cross-lingual textual entailment, and propose an effective architecture to improve its performance. To this end, we propose a novel transfer learning mechanism from a certain language pair to another language pair, equipping the latter with knowledge and features obtained from the former language pair efficiently. We also introduce a joint loss objective accompanied by a traditional cross-entropy loss function. We perform a wide range of experiments and show that our baseline model, while equipped with the novel transfer learning technique, significantly improves the performance of CLTE tasks. We evaluate our approaches on a standard English - Hindi cross-lingual dataset (Saikh et al., 2019), derived from the Stanford Natural Language Inference (SNLI) (Bowman, Samuel R. and Angeli, Gabor and Potts, Christopher, and Manning, Christopher D., 2015) Corpus. To further assess the effectiveness of our proposed methodology across a wide range of languages, we also perform a wide range of experiments on four languages of the XNLI (Conneau et al., 2018) dataset i.e *French*, *German*, *Bulgarian* and *Turkish* for Cross-lingual natural language inference tasks and show improvement of results us-

*Work done while he was an intern at IIT Patna.

ing our proposed approach (c.f. Section 3.2) than the existing baseline (c.f. Section 3.1). Furthermore, we provide a detailed analysis of the effects of employing our proposed transfer learning mechanism on the hidden state representations of the models used. Though we analyze our transfer learning mechanism on BERT (Devlin et al., 2019), this transfer learning technique is architecture-independent, and it can be used with both Transformer (Vaswani et al., 2017) based and Recurrent Neural Network-based models.

This paper mainly focuses on Cross-lingual Textual Entailment (CLTE) and methods to improve the performance of the existing models in CLTE. The specific attributes of the current work are summarized as follows:

1. We introduce a novel transfer learning mechanism from one language pair to another for better performance in the textual entailment problem in a cross-lingual setting.
2. To enable the transfer learning, we introduce a novel joint loss function along with the traditional cross-entropy loss function and show that introducing this mechanism along with the joint loss function improves the performance of our baseline model significantly.
3. To establish the effectiveness and robustness of our proposed transfer learning mechanism and joint loss function, we perform a wide range of experiments on a standard English - Hindi CLTE dataset (Saikh et al., 2019) and on chosen language pairs (*viz.* language pairs from French, Turkish, German, Bulgarian) from the `XNLI` dataset (Conneau et al., 2018). We show that our proposed method performs consistently better than the baseline used, sometimes surpassing the current state-of-the-art. We perform a detailed comparison of our proposed method and the current state-of-the-art for CLTE.
4. Lastly, we also perform a detailed analysis of the transfer learning mechanism and its effect on the hidden state representations by applying this mechanism. We establish that our proposed methods are indeed efficient in transferring features from one language pair to another language pair, which ultimately improves the performance.
5. Our proposed method uses the standard *Multilingual BERT-base* having 110M parameters and in some cases it outperforms the current state-of-the-art (SOTA) model i.e., *XLM-R-base* (Conneau et al., 2020), which contains 270M parameters. This makes our method more parameter efficient than the existing SOTA without losing much performance.

2. Related Work

Natural Language Inference is a widely-studied problem. In a monolingual setting, the study of Natural Language Inference (NLI) has been prevalent, especially in the English language. Our focus is on improving cross-lingual NLI, i.e., NLI, where the sentences are in two different languages.

Textual Entailment: One of the first notable advancement of Natural Language Inference is the creation of FraCaS (Framework for Computational Semantics) dataset (Cooper et al., 1996) which was created to evaluate the semantics of Textual Entailment. The dataset was divided into three classes: Entailment, Contradiction, and Neutral. After several years, in PASCAL RTE (2005) challenge (Dagan et al., 2006) the NLI dataset contained real-life premise-hypothesis pairs. One of the most notable works in the field of NLI is the surfacing of SNLI (Bowman, Samuel R. and Angeli, Gabor and Potts, Christopher, and Manning, Christopher D., 2015), and MultiNLI (Williams et al., 2018) datasets. Many models have been built on these datasets *viz.* (Chen et al., 2017; Rocktäschel et al., 2016; Parikh et al., 2016). Recently, Google T5 (Raffel et al., 2019) model has obtained state-of-the-art in many NLP tasks, including monolingual NLI.

Cross-lingual Entailment: Cross-Lingual Natural Language Inference is an extension to the NLI task. (Mehdad et al., 2010) introduced this concept of CLTE. Later, SemEval-2012 (Negri et al., 2012) and SemEval-2013 (Negri et al., 2013) with CLTE's application scenario in content synchronization were organized. Recently, multilingual research in NLP has seen much interest, especially in the Deep Learning era. (Conneau et al., 2018) introduced XNLI dataset containing 14 languages apart from English. There have been several attempts to perform cross-lingual NLI (Aghajanyan et al., 2021; Le et al., 2020; Cui et al., 2021; Xue et al., 2022; Conneau et al., 2020). However, cross-lingual NLI on low-resource languages remains mostly unexplored. This paper proposes a novel transfer-learning method and compares our findings with the state-of-the-art model, XLM-R.

3. Methodology

We define a *baseline* (c.f. 3.1), which is Multilingual Cased BERT Base (Devlin et al., 2019) architecture with 12 hidden layers. We fine-tune this baseline model to perform the CLTE task. To improve the baseline model, we introduce a novel fine-tuning procedure where we propose a joint training loss function consisting of an additional loss term and a standard cross-entropy loss term associated with supervised classification tasks. We refer this improved baseline as *BERT-KLD*, defined in Section 3.2.

3.1. Baseline Model

We use the Multilingual Cased BERT Base[1] model as our baseline, and it is described below in details.

Input Layer: We input a sentence pair consisting of premise and hypothesis to the model. It is important that we use the premise and hypothesis already translated

[1] `https://github.com/google-research/bert/blob/master/multilingual.md`

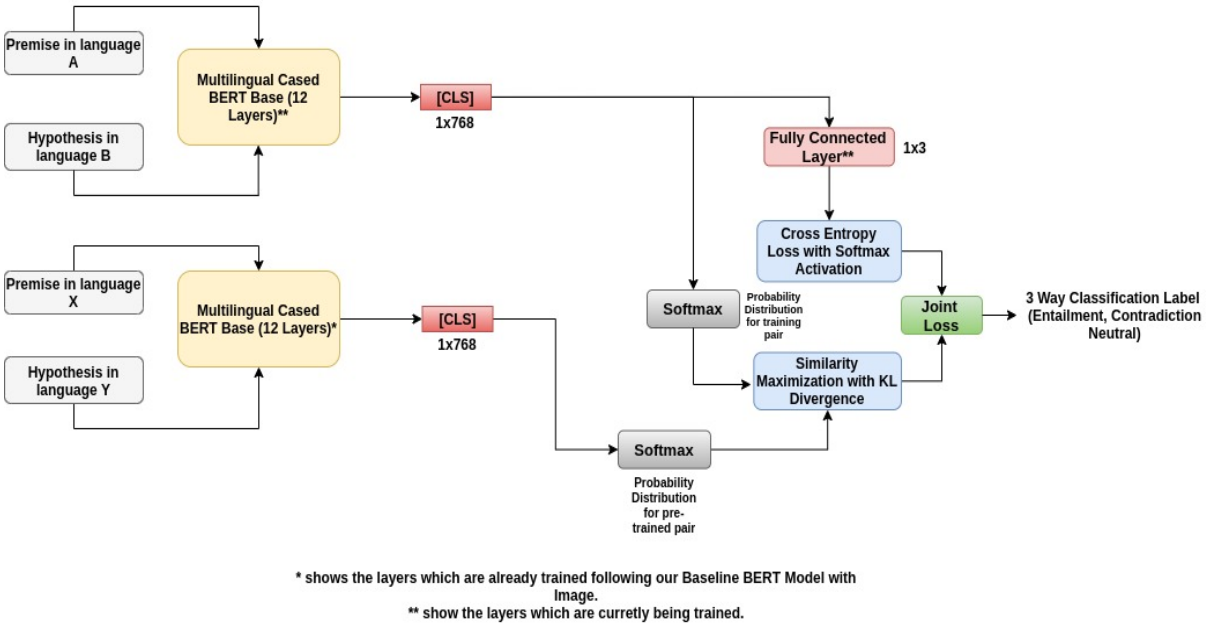

Figure 1: Schematic of transfer learning method to train *baseline* architecture

into two different languages.

Hidden Layers: After the inputs are fed into the model, we take the final hidden state (i.e., the output of the Transformer) for the first token, which corresponds to the special *[CLS]* token embedding. We obtain this context vector (we call it C) with a dimension of $1 \times H$ (in this case of BERT base, $H = 768$).

Classification Layer: The hidden representation vector C is then fed into a feed-forward layer (*FFN*) (with $K \times H$ dimensional weight, denoted by W), where K is the number of ground truth labels i.e Entailment, Neutral, Contradiction (in this case, K = 3). We call this intermediate representation as F.

$$F = C \cdot W^T + b \tag{1}$$

where $\cdot$ denotes the matrix multiplication between the weight matrix, W and the context vector C and b is a bias term. The label probabilities are computed with a standard softmax, $P = softmax(F)$. P has a dimension of $1 \times K$. All the BERT and feed-forward layer parameters are fine-tuned jointly to maximize the log-probability of the correct label by employing standard cross-entropy loss.

3.2. Improved Baseline Model (BERT-KLD)

To further improve the performance of our baseline model, we introduce a novel loss formulation that facilitates transfer learning from an *already trained baseline model* ***(Teacher Model; henceforth called P)*** in language pair X-Y to a *baseline model we are training* ***(Student Model; henceforth called Q)*** in language pair A-B. We name this Student model as **BERT-KLD**. Firstly, we train our baseline model denoted as P in the X-Y language pair (X is the language of premise, Y is the language of hypothesis). After training P, we initialize a baseline model Q, which we train using a novel loss formulation. We use A-B language pair (A is the language of premise, B is the language of hypothesis) as input to Q.

For a given input example consisting of a premise and a hypothesis, we translate them to X-Y and A-B language pairs, respectively. We then input the translated sentence pairs (X-Y) and (A-B) to P and Q, respectively. We denote *hidden1* and *hidden2* as the final hidden state i.e. output of Transformer for the first token, which corresponds to the special *[CLS]* token embedding for the models P and Q, respectively. We optimize the parameters of model Q by employing the following loss function.

$$JointLoss = CrossEntropy(a, b) + \lambda \times D_{\mathrm{KL}}(h1 \parallel h2) \tag{2}$$

Where,

- a is log probability of the associated label output by model Q and b is one hot representation of the ground truth label.
- *h1*, *h2* are the softmax probability distribution for *hidden1* and *hidden2* respectively.

$$h1 = softmax(hidden1) \tag{3}$$

$$h2 = softmax(hidden2) \tag{4}$$

- $D_{\mathrm{KL}}(P \parallel Q)$ denotes the KL-Divergence between two probability distributions (P(x) and Q(x)) defined as,

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{5}$$

- λ refers to a hyper-parameter associated with KL-Divergence loss.

A detailed training algorithm is described below (Algorithm 1).

Algorithm 1 Algorithm to fine-tune *BERT-KLD* for CLTE

Require: : A baseline model(**P**) already fine-tuned on language pair **X** and **Y**.
Require: : Language pair **A** and **B** to be used in training a new BERT model(**Q**) with parameters θ.
Require: : Hyper-parameter λ initialized between 0.01 and 0.5.
while not converges **do**
  Sample a batch(with Batch_Size = B) of premise and hypothesis as a set of sentence pairs τ.
  Verify the premise and hypothesis for each sentence pair is translated into language pairs **X** and **A**. *and* **Y** and **B** respectively.
  Batch Loss $\leftarrow 0$
  for each α in τ **do** ▷ α is a sentence pair.
    Construct *one_hot_labels* as one hot representation of ground truth label associated with α
    $N \leftarrow no\_of\_labels$
    Construct an example **I1** for already trained teacher model (P) using premise and hypothesis from α, translated into language **X** and **Y**.
    Construct an example **I2** for to be trained student model (Q) using premise and hypothesis from α, translated into language **A** and **B**.
    $h1 \leftarrow hidden\_representation\_from\_P(\mathbf{I1})$ ▷ $dim(\mathbf{h1}) \leftarrow$ **768*1**
    $h2 \leftarrow hidden\_representation\_from\_Q(\mathbf{I2})$ ▷ $dim(\mathbf{h2}) \leftarrow$ **768*1**
    $logits \leftarrow Feed\_Forward\_Network(\mathbf{h2})$
    $log\_probs \leftarrow log(softmax(logits))$
    Cross-Entropy Loss $\leftarrow -\sum_{n=1}^{N} log\_probs * one\_hot\_labels$
    KLD Loss $\leftarrow D_{\mathrm{KL}}(softmax(\mathbf{h1}) \parallel softmax(\mathbf{h2}))$
    Total Loss $\leftarrow$ *Cross-Entropy Loss* $+ \lambda *$ *KLD Loss*
    Batch Loss $\leftarrow$ *Batch Loss* $+$ *Total Loss*
  end for
  Loss $\leftarrow (1/Batch\_Size) * \sum_{i=1}^{Batch\_Size}$ *Batch Loss[i]*
  Optimize Parameters θ of model Q with standard BERT optimization technique with loss computed in the previous line.
end while

3.3. Intuitive Explanation of Adjoin Loss Function

While for multiclass classification tasks, the trivial Cross-Entropy loss function is used, in this paper, we introduce a better way to work with CLTE tasks (*a multiclass classification task*) by using a weighted sum of two different loss terms i.e. the trivial cross-entropy loss and the Kullback-Leibler Divergence or KL-Divergence (c.f. equation 2). According to the setting described in Section 3.2, we obtain the hidden representation $hidden1$, $hidden2$ from model P and Q, respectively. Next, we convert them into *softmax* probabilities $h1$, $h2$ and compute the KL-Divergence between $h1$ and $h2$ as shown in Equation 6.

$$D_{\mathrm{KL}}(h1 \parallel h2) = \sum_{x \in \mathcal{X}} h1 \log\left(\frac{h1}{h2}\right) \quad (6)$$

Thus, while training the second model Q, the term $D_{\mathrm{KL}}(h1 \parallel h2)$ captures all the important learned representations of the already trained model *P* from $h1$ and transfers them to $h2$, thereby enriching information that is useful for Textual Entailment problem employing Q in A-B language-pair. Next, we compute $JointLoss$ defined in Equation 2. Later we discuss the consistent improvements achieved using this adjoin loss function over the trivial Cross-Entropy loss function. We demonstrate training process of *BERT-KLD* in schematic 1.

4. Datasets and Experimental Setup

4.1. Datasets

We perform experiments on four European languages, i.e. *French (**fr**)*, *German (**de**)*, *Bulgarian (**bg**)* and *Turkish (**tr**)* of `XNLI` (Conneau et al., 2018) dataset for the *baseline* and the proposed *BERT-KLD* model (i.e. improvement on baseline using adjoin loss). We use all the $5,000$ sentence pairs from the `XNLI-Test` set for each language and split them into train and test set, respectively, with 10% of sentences in our test set. For validation, we have used 5% examples of the training set. To maintain the alignment of training and test set between each of the language pairs, we use a fixed random seed across every language pairs.

To assess the performance of our models, we further test our baseline and modified architecture on a dataset (Saikh et al., 2020) for cross-lingual NLI on English-Hindi language pairs. This dataset is derived from the SNLI (Bowman, Samuel R. and Angeli, Gabor and Potts, Christopher, and Manning, Christopher D., 2015) Corpus. To differentiate this dataset from `XNLI` dataset, we call this dataset as `EH-XNLI`. We report the performance of our proposed models on both `XNLI` & `EH-XNLI` datasets in Section 5 below. For EH-XNLI, we have used the same train-dev-test split as in (Saikh et al., 2020).

4.2. Experimental Setup

For training of both *baseline* and *BERT-KLD*, we use the following training setup: We employ a batch size of 28 and an initial learning rate of $2e-5$ and a maximum sequence length 128. We use 42 as a random seed. A warm-up proportion of 0.1 is used, where warm-up is a period where the learning rate is low and gradually increases, usually helping the training procedure. For all experiments, we run each of them for 5 epochs. Both training procedures are executed on NVIDIA Tesla P100 GPU environment. TensorFlow framework is used to perform all the experiments.

5. Results and Discussions

5.1. Evaluation on EH-XNLI dataset

5.1.1. Evaluation using baseline model

Firstly we evaluate our baseline model (c.f. Section 3.1) on English-Hindi and Hindi-English language pairs from our `EH-XNLI` dataset. From the Table 1, we see that the accuracy obtained are 68.21% and 71.08%, respectively, whereas the reported accuracy scores according to (Saikh et al., 2020) are 69% and 72%, respectively.

5.1.2. Evaluation using BERT-KLD

As discussed in Section 3.2, while training model Q for A-B language pair using our proposed adjoin training loss, P is an already trained baseline model on language pair X-Y. Note that, when A-B is English-Hindi, then X-Y is Hindi-English and vice-versa. From Table 1,

for both English-Hindi and Hindi-English language pairs, we obtain 2.15% (*from* 68.21% *to* 70.36%) and 1.22% (*from* 71.08% *to* 72.30%) improved accuracy than the baseline model, indicating positive transfer of knowledge from our baseline model trained on English-Hindi (or Hindi-English) language pairs to a model which we train on Hindi-English (or English-Hindi) language pairs, equipped with adjoin loss function (c.f. Equation 2).

Model	Premise	Hypothesis	Accuracy (%)
Baseline	Eng	Hin	68.21
	Hin	Eng	71.08
BERT-KLD	Eng	Hin	70.36
	Hin	Eng	72.30

Table 1: Results obtained on the baseline and the improvement on baseline on different input modalities. Here, **Eng**: English; **Hin**: Hindi; **KLD**: Kullback–Leibler divergence Loss Function. *For all the experiments, λ=0.2*

5.2. Evaluation on XNLI dataset

5.2.1. Evaluation using baseline

We perform 6 experiments using our baseline model (c.f. 3.1) with language pairs *fr-de*, *fr-tr*, *fr-bg*, *de-tr*, *de-bg*, *tr-bg*. Of all these experiments, we obtain the highest accuracy in *fr-de* (57.6%) and lowest in *tr-bg* (48%). The results of these experiments are tabulated in Table 2. We specifically choose low-resource language like Bulgarian (bg) and Turkish (tr) to show the effects of the proposed transfer learning method on low-resource languages.

	Language Pairs					
	fr-de	*fr-tr*	*fr-bg*	*de-tr*	*de-bg*	*tr-bg*
Accuracy	57.6	50	56	51.6	54.8	48

Table 2: Results obtained in terms of accuracy (%) on four languages of XNLI dataset using *baseline*

5.2.2. Evaluation using BERT-KLD

We conduct experiments using *BERT-KLD* on all of the language pairs and the results of the experiment is tabulated in the Table 3. In Table 3, each cell shows accuracy score of the *BERT-KLD* model trained using the language pair shown in respective *column* by using hidden representation from the *baseline* model (thereby optimizing adjoin loss) already trained in language pair shown in respective *row*. For example, training *BERT-KLD* on *de-bg* by obtaining hidden representation from model *baseline* already trained on *de-tr* leads to an accuracy of 58.4%. For each column of Table 3, we highlight the best accuracy score.

Comparing Tables 2 & 3, we observe that our improved baseline with adjoin loss i.e. *BERT-KLD* performs consistently well over the baseline. For example, *tr-bg* baseline accuracy is 48% and the accuracy increases to 58.2% when we train *BERT-KLD* model in *tr-bg* (with hidden representation from baseline model already trained in *de-tr*) using our novel training algorithm and select optimal hyper-parameter λ = 0.1 by grid search. Thus, we obtain 10.2% improvement in accuracy for *tr-bg* language pair.

	Transferred					
Transferring	*fr-de*	*fr-tr*	*fr-bg*	*de-tr*	*de-bg*	*tr-bg*
fr-de	-	50.4	57	50	57	51
fr-tr	59	-	57	51	**58**	54.2
fr-bg	**59.2**	**52**	-	51	57.6	53.8
de-tr	58.6	49.8	**58.4**	-	55.4	**58.2**
de-bg	57.8	51.8	56.8	49.4	-	54
tr-bg	58.8	51	57.2	49.6	57.2	-

Table 3: Results obtained in terms of accuracy (%) on four languages of XNLI dataset using **BERT-KLD**. Here, **X-axis**: Knowledge is being transferred from these language pairs (Rows); **Y-axis**: Knowledge is being transferred to these language pairs (Columns); **Language identifier:** French: (fr), German: (de), Bulgarian: (bg), Turkish (tr).

5.2.3. Hyper-parameter Optimization

As stated in Algorithm 1, we introduce a hyper-parameter associated with the KL-Divergence loss to account for its importance. We conduct a grid search to optimize hyper-parameter λ for each of the experiments and find its optimal value ranging between 0.01 to 0.5 depending on the language pair. For hyper-parameter optimization, we have used 5% of the XNLI train set, which we have formed for our experiments 4.1. For each of the experiments in Table 3, we use the optimal hyper-parameter value that was found by the grid search. For finding optimal λ in our XNLI experiments, we use grid search and ran the experiments for the following values 0.01, 0.05, 0.1,0.2, 0.25, 0.27, 0.3, 0.4, 0.5. For each of the XNLI experiment mentioned in Table 3, the optimal values of hyper-parameter λ used in the experiments are shown in the Table 4

	Transferred					
Transferring	*fr-de*	*fr-tr*	*fr-bg*	*de-tr*	*de-bg*	*tr-bg*
fr-de	-	0.2	0.2	0.2	0.2	0.2
fr-tr	0.1	-	0.1	0.5	0.2	0.2
fr-bg	0.1	0.1	-	0.2	0.5	0.27
de-tr	0.01	0.1	0.5	-	0.3	0.1
de-bg	0.27	0.1	0.5	0.2	-	0.1
tr-bg	0.1	0.1	0.1	0.25	0.3	-

Table 4: Optimal λ used for each XNLI experiment

5.2.4. Comparison to the State-of-the-arts

We use *XLM-R* model, which is currently the state-of-the-art (SOTA) model for learning cross-lingual representations on XNLI dataset. To compare this model's performance to *BERT-KLD*, we train *XLM-R* model on

the same XNLI training set that we used to train our *baseline* model and *BERT-KLD* both. To make fare comparisons between these models, we use *XLM-R-base* model and train it using the same experimental setup used to train our proposed model. It is important that during training, the *XLM-R* model was trained completely in English and later tested in 6 language pairs (*fr-de*, *fr-tr*, *fr-bg*, *de-tr*, *de-bg*, *tr-bg*). The comparison with our best performing model (BERT-KLD) are tabulated in Table 5. It is noteworthy that in some of the language pairs (*fr-de*, *tr-bg*), our model exceeds the performance of the SOTA model while in some of the language pairs (*fr-tr*, *fr-bg*, *de-tr*) the performance is very close to SOTA. We also note that our proposed model only has 110M parameter count, which is far lesser than the state-of-the-art model (XLM-R) which has 270M parameters. Despite having such lower no of parameters, our model is comparable to the state-of-the-art model.

	Language Pairs					
Models	*fr-de*	*fr-tr*	*fr-bg*	*de-tr*	*de-bg*	*tr-bg*
XLM-R (SOTA)	58.6	**53.2**	**58.8**	**52.4**	**64.6**	57.8
BERT-KLD (Ours)	**59.2**	52	58.4	51	58	**58.2**

Table 5: Comparison of BERT-KLD with State-of-the-art. For BERT-KLD, best accuracy scores are listed from Table 3

5.3. Comparison to Knowledge Distillation

Our transfer learning objective resembles Knowledge Distillation (KD) Framework (Hinton et al., 2015). We have one fundamental difference with the KD framework concerning the transfer objective. In the classical KD framework, the objective is to minimize the KL-Divergence between the *soft logit predictions* of the Teacher and Student model. In contrast, we try to minimize the KL-Divergence between the *hidden state* of the Teacher and Student. To compare our method with the classical KD framework, we apply KD to improve the performance of the baseline model (*Student*) on *fr-tr* language pair with the help of another baseline (*Teacher*) which is already trained on *fr-bg* pair. Figure 3 shows the result with different temperature values, and the highest obtained accuracy using the KD framework is 50.8%. From Table 3, using the same training setup, our method achieves 52% accuracy, which is more than what is achieved using the classical KD framework. Performance of *baseline* model for *fr-tr* language pair is 50%, which is only slightly increased by employing Classical Knowledge distillation. Hence, employing KD for CLTE might be beneficial for CLTE tasks rather than simply using a baseline model.

5.4. Analysis of Hidden State Dynamics

To examine the effect of hidden state in model performance, we perform empirical analysis to show how our transfer learning objective equips *BERT-KLD* model to perform better than our baseline. We choose *de-bg* language pair for our analysis. We train a baseline model on *de-bg* language pair, which obtains an accuracy of 54.8%. Using our training objective, the updated baseline model (*BERT-KLD*; the Student Model) achieves an accuracy of 55.4%. This is done by transfer learning of hidden states from a baseline model (the Teacher Model) trained on *de-tr* language pair.

The training dynamics are clearly seen in Figure 2. On the left figure, we plot the t-SNE projection (van der Maaten and Hinton, 2008) of the hidden state both for trained Teacher model, trained on *de-tr* pair and student model, trained on *de-bg*. They are trained independently of each other, and there is no overlap between the hidden state as expected. The labels *TRSW* denote the instances where the teacher model classifies correctly, but the student model fails, and *TRSR* denotes the instances where both teacher and student classify correctly. Overlap between *TRSR* and *TRSW* instances would indicate that the teacher model influences the student model to classify more instances correctly, which has already been predicted correctly by the teacher model. So, intuitively we expect our transfer learning objective to minimize the distance between *TRSR* and *TRSW* instances. This can be clearly seen from Figure 2. We can see the closeness between *TRSR* and *TRSW* pairs after transfer learning as compared to no overlap before applying the transfer learning method.

Timeline	d_TRSR	d_TRSW	num_TRSR	num_TRSW
BT	15.4	16.9	80	178
AT	13.2	16.3	86	172

Table 6: d_TRSR & d_TRSW denotes average euclidean distance between Teacher-Student TRSR and TRSW instances on 768-dimensional vector space. num_TRSR and num_TRSW denotes the number of those instances respectively. **BT:** Before Transfer. **AT:** After Transfer.

From Table 6, it can be seen that the average Euclidean distances between Teacher-Student *TRSR* instances have decreased after transfer learning. The cases that are correctly classified by both teacher and student models (TRSR instances) are more closely situated after transfer learning than before the transfer. This indicates the positive transfer of knowledge from Teacher to Student. Moreover, the number of *TRSR* instances has increased, indicating that both teacher and student model agrees with their prediction of more correct examples.

Note that though we analyze this transfer learning methodology for two specific language pairs (*de-tr* & *de-bg*), a similar kind of analysis can be performed on other pairs of language pairs to assess the effectiveness of transfer learning. This kind of reasoning is also applicable for the comparison of our model with the classical KD framework in Section 5.3, where *fr-bg* and *fr-tr* language pairs are specifically used to demonstrate the effectiveness of our training method when compared to classical Knowledge distillation.

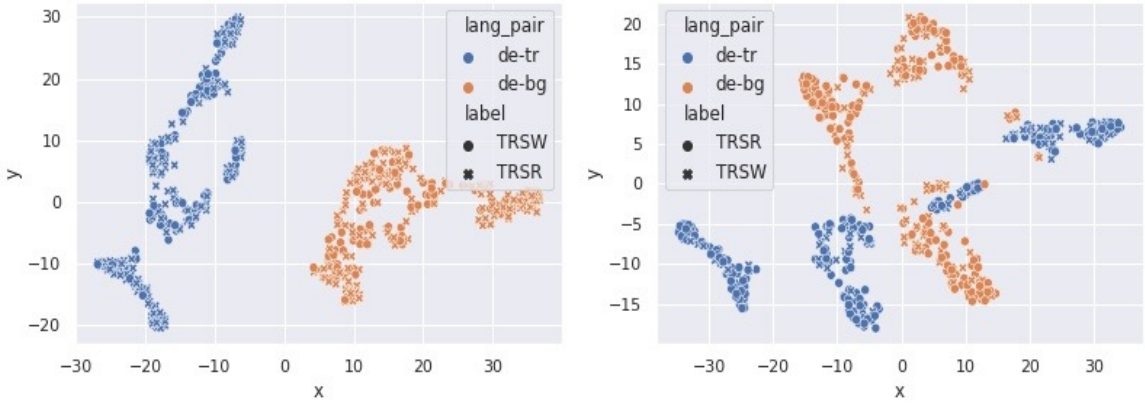

Figure 2: **Left:** t-SNE projection of the hidden state vector of a trained baseline model before transfer on *de-tr* and *de-bg* language pairs. **Right:** t-SNE projection of hidden state vector of both Teacher (P) and Student (Q) model after transfer learning. Before transfer took place (in the *left*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

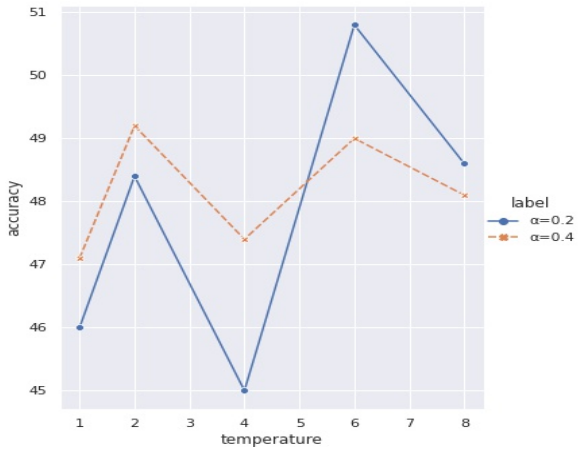

Figure 3: Performance of *BERT-KLD* employing classical KD framework. Transferring and transferred language pairs are *fr-bg* and *fr-tr* respectively. Peak accuracy obtained is 50.8%. α denotes the hyper-parameter associated with the extra loss term used to train the Student model. It is seen varying temperature parameter in softmax distribution can result in widely increased/decreased performance, with higher values of temperature usually giving better performance.

6. Conclusion and Future Work

In this paper, we present a novel transfer learning algorithm to tackle CLTE tasks, which yields improved performance over the existing baseline. We show the robustness of our proposed algorithm by conducting experiments on four different European languages facilitating consistent improvement in the performance, even surpassing SOTA for some language pairs. It is also worth noting that our proposed training algorithm is model agnostic in the sense that it can be used as a tool for transfer learning in CLTE employing other deep learning based models used in NLP (e.g., LSTM (Hochreiter and Schmidhuber, 1997)). In the future, we would like to extend our work by incorporating multimodal features (e.g., 'acoustic' that consists of pitch, voice quality and visual information like video frame which captures gesture and posture) in cross-lingual NLI setting, incorporating novel textual and visual information fusion techniques and building a theoretical ground for our novel training algorithm.

7. Acknowledgements

The authors gratefully acknowledge the support from Visvesvaraya Young Faculty Research Fellowship (YFRF) Grant of Ministry of Electronics and Information Technology, Govt of India.

8. Bibliographical References

Aghajanyan, A., Shrivastava, A., Gupta, A., Goyal, N., Zettlemoyer, L., and Gupta, S. (2021). Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Bowman, Samuel R. and Angeli, Gabor and Potts, Christopher, and Manning, Christopher D. (2015). A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal. Association for Computational Linguistics.

Chen, Q., Zhu, X., Ling, Z.-H., Wei, S., Jiang, H., and Inkpen, D. (2017). Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668,

Vancouver, Canada, July. Association for Computational Linguistics.

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.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, et al., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Cooper, R., Crouch, D., Eijck, J. V., Fox, C., Genabith, J. V., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., Pulman, S., Briscoe, T., Maier, H., and Konrad, K. (1996). Using the Framework.

Cui, Y., Che, W., Liu, T., Qin, B., and Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3504–3514, jan.

Dagan, I., Glickman, O., and Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, pages 177–190, Southampton, UK. Springer-Verlag.

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.

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

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May. European Language Resources Association.

Mehdad, Y., Negri, M., and Federico, M. (2010). Towards Cross-lingual Textual Entailment. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 321–324, Los Angeles, California.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., and Giampiccolo, D. (2012). SemEval-2012 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 399–407, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Negri, M., Marchetti, A., Mehdad, Y., Bentivogli, L., and Giampiccolo, D. (2013). SemEval-2013 Task 8: Cross-lingual Textual Entailment for Content Synchronization. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 25–33, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. (2016). A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kociský, T., and Blunsom, P. (2016). Reasoning about Entailment with Neural Attention. In Yoshua Bengio et al., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Saikh, T., Anand, A., Ekbal, A., and Bhattacharyya, P. (2019). A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features. In *International Conference on Applications of Natural Language to Information Systems*, pages 345–358. Springer.

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.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polo-

sukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Williams, A., Nangia, N., and Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., and Raffel, C. (2022). ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 03.