CL-NERIL: A Cross-Lingual Model for NER in Indian Languages
(Student Abstract)

Akshara Prabhakar¹, Gouri Sankar Majumder², Ashish Anand²
¹National Institute of Technology Karnataka, Surathkal
²Indian Institute of Technology Guwahati
akashblr555@gmail.com, gourisankar@iitg.ac.in, anand.ashish@iitg.ac.in

Abstract
Developing Named Entity Recognition (NER) systems for Indian languages has been a long-standing challenge, mainly owing to the requirement of a large amount of annotated clean training instances. This paper proposes an end-to-end framework for NER for Indian languages in a low-resource setting by exploiting parallel corpora of English and Indian languages and an English NER dataset. The proposed framework includes an annotation projection method that combines word alignment score and NER tag prediction confidence score on source language (English) data to generate weakly labeled data in a target Indian language. We employ a variant of the Teacher-Student model and optimize it jointly on the pseudo labels of the Teacher model and predictions on the generated weakly labeled data. We also present manually annotated test sets for three Indian languages: Hindi, Bengali, and Gujarati. We evaluate the performance of the proposed framework on the test sets of the three Indian languages. Empirical results show a minimum 10% performance improvement compared to the zero-shot transfer learning model on all languages. This indicates that weakly labeled data generated using the proposed annotation projection method in target Indian languages can complement well-annotated source language data to enhance performance. Our code is publicly available at https://github.com/aksh555/CL-NERIL.

Proposed Approach

Weakly Labeled Data Generation: We propose an annotation-projection-based approach to create weakly labeled data from a large bi-lingual parallel corpus (Ramesh et al. 2021). Let \( \{(S_i, T_i)\}_{i=1}^{N} \) having \( N \) sentences be a parallel corpus of English and a low-resource language, where \( (S_i, T_i) \) represents a pair of sentences which are translations of each other. We first generate the named entity tag-sequence \( y_i^e \) for the source (English) sentence \( S_i \) using Flair NER trained on CoNLL dataset and store the confidence score of each detected entity as \( NER \) score. Next, we determine word and subword alignments between \( S_i \) and \( T_i \). We use XLM-R (Conneau et al. 2020) for word-level alignment. We align two words that are the most similar in terms of normalized cosine similarity. Ties are resolved by giving preference to aligned smaller index word in the sentence. This is a local optimal solution, and only mutual alignments are identified, resulting in many entity words getting missed. Since our focus is to project the entity tags to the low resource language, we try to further align the missed entity words (if any) by using the Match algorithm (Jalili Sabet et al. 2020). Here, we utilize the subword-level alignments with mBERT (Devlin et al. 2019) embeddings. This process ensures a near alignment for all entity words identified in the source side to a word(s) on the target side. The normalized cosine similarity that we obtain via these alignments for every word is the alignment score. In case a word gets aligned to many words, we propagate the source tag label \( y_i^e \) to target sentence according to the alignments to get \( y_i^t \). Finally, we select the top 40% (empirically chosen value) sentences based on the
sentence score to filter the good quality sentences as weakly labeled data. Sentence score for a sentence with \( k \) number of entities is calculated as
\[
\text{score} = \frac{1}{k} \sum_{i=1}^{k} \log(\text{alignment score} \times \text{NER score})
\]

**Teacher-Student Model:** We employ Teacher-Student model to transfer knowledge from source to target language. Firstly, for the Teacher model, we fine-tune mBERT by training it on CoNLL English data. The Student model is initialized with pre-trained mBERT layers. Next, our weakly labeled sentences are passed to both the Teacher model and Student model to generate pseudo labels \( y_{\text{teach}}, y_{\text{stud}} \) respectively, i.e., probability distributions over the tag set for each word \( w_i \) in the target sentence. Finally, the Student model is trained by optimizing the mean squared error (MSE) loss between the two predicted distributions \( y_{\text{teach}}, y_{\text{stud}} \) (considering the Teacher model’s predicted labels as soft), and negative log-likelihood (NLL) of the generated weak label of each word \( y_i \) and the predicted labels \( y_{\text{stud}}(y_{\text{teach}}, y_i) \). Model parameter details and hyperparameter tuning are discussed in the supplementary material.

\[
\text{Loss} = \frac{1}{N} \sum_{i=1}^{N} \{ \text{MSE}(y_{\text{teach}}, y_{\text{stud}}) + \text{NLL}(y_{\text{stud}}, y_{\text{teach}}) \}
\]

**Results & Analysis**

We evaluate CL-NERIL on manually annotated test sets having at least 1000 sentences for three languages – Bengali (bn), Gujarati (gu), and Hindi (hi). We compared our model with three state-of-the-art models for cross-lingual NER. Table 1 summarizes the results. Wu and Dredze (2019) used mBERT model for zero-shot transfer (which we use as the Teacher model in our approach) to train NER on English and showed cross-lingual transferability of the model on different target languages. We compared the self-learning model, BOND (Liang et al. 2020), which is one of the strongest baselines on the weakly labeled data (generated by our projection approach). Teacher Student model (Wu et al. 2020) uses knowledge distillation method (Sanh et al. 2020) to transfer knowledge of the NER model trained on English to the target language and showed improvement over the zero-shot transfer. To assess the quality of our generated weak labels by annotation-projection approach, we trained BERT based monolingual NER model on the weakly labeled data. Its performance corroborated the quality of our generated weakly labeled data. Our model, CL-NERIL, due to joint optimization of the MSE loss with teacher prediction and NLL loss on weak labels, does a better job in cross-lingual transfer and showed significant improvement over the Teacher Student model on all languages, and even over BOND for two languages.

**Conclusion & Future Work**

In this work, CL-NERIL, an end-to-end cross-lingual model for NER task on Indian languages in low-resource settings was presented. It leverages the easily procurable weakly labeled data in the target language to complement the gold standard data in the source language to enhance the performance on target languages. We observe that there are two sources of noise in this approach – from alignment and NER on the source side. We are currently working on a novel noise-aware loss function to take care of these noise sources.

**Table 1: Benchmarking CL-NERIL against State-of-the-art.**

<table>
<thead>
<tr>
<th>Model</th>
<th>bn</th>
<th>gu</th>
<th>hi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero-shot (Wu and Dredze 2019)</td>
<td>50.67</td>
<td>56.59</td>
<td>70.28</td>
</tr>
<tr>
<td>BOND (Liang et al. 2020)</td>
<td>64.19</td>
<td>79.28</td>
<td><strong>80.14</strong></td>
</tr>
<tr>
<td>Teacher-Student (Wu et al. 2020)</td>
<td>53.79</td>
<td>58.80</td>
<td>72.37</td>
</tr>
<tr>
<td><strong>Monolingual on Weak Data (Ours)</strong></td>
<td><strong>71.21</strong></td>
<td><strong>80.39</strong></td>
<td><strong>79.80</strong></td>
</tr>
<tr>
<td><strong>CL-NERIL (Ours)</strong></td>
<td><strong>73.34</strong></td>
<td><strong>80.73</strong></td>
<td><strong>79.69</strong></td>
</tr>
</tbody>
</table>

**References**


