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

Akshara Prabhakar[1], **Gouri Sankar Majumder**[2], **Ashish Anand**[2]

[1]National Institute of Technology Karnataka, Surathkal
[2]Indian Institute of Technology Guwahati
akshblr555@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.

Introduction

Named Entity Recognition (NER), a fundamental task in natural language processing, deals with detecting named entities and their classification into certain predefined categories. However, the performance of NER models depends on the amount and quality of annotated data available. Indian languages, for one, lack such accurate entity annotated corpora to build and benchmark NER systems. Cross-lingual NER attempts to address this challenge by transferring knowledge from a high-resource source language having abundant labeled entities to a low-resource target language having few or no labels. The projection-based approaches (Mayhew, Tsai, and Roth 2017; Xie et al. 2018) applying word translations perform poorly primarily due to syntactic word order differences between English and Indian languages. Problems arising due to word order differences can be mitigated using word alignment method. This work uses multilingual embedding-based word alignment method to propose an annotation projection method to generate weakly labeled data in the target language. The annotation projection method is integrated into an end-to-end cross-lingual transfer learning framework for NER in Indian languages, referred to as CL-NERIL. Primary contributions of this paper are : (1) An end-to-end framework, CL-NERIL, with a novel annotation projection method and joint optimization, (2) Clean test sets of at least 1000 sentences for the three Indian languages, (3) CL-NERIL obtains at least 10% improvement compared to zero-shot baseline models.

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^s 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 assign the same tag to all words, and to resolve multiple tag matches, we take the one giving the highest score. Then, we propagate the source tag label y_i^s 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

$$score = \frac{1}{k}\sum_{i=1}^{k}\log(\text{alignment score} * \text{NER score})$$

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^{teach}, y^{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^{teach}, y^{stud} (considering the Teacher model's predicted labels as soft), and negative log-likelihood (NLL) of the generated weak label of each word (y^t) and the predicted labels (y^{stud}). Model parameter details and hyperparameter tuning are discussed in the supplementary material.

$$Loss = \frac{1}{N}\sum_{i=1}^{N}\{MSE(y_i^{teach}, y_i^{stud}) + NLL(y_i^{stud}, y_i^{t})\}$$

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.

Model	**bn**	**gu**	**hi**
Zero-shot (Wu and Dredze 2019)	50.67	56.59	70.28
BOND (Liang et al. 2020)	64.19	79.28	**80.14**
Teacher-Student (Wu et al. 2020)	53.79	58.80	72.37
Monolingual on Weak Data (Ours)	71.21	80.39	79.80
CL-NERIL (Ours)	**73.34**	**80.73**	79.69

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

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.

References

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 *ACL*, 8440–8451.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186.

Jalili Sabet, M.; Dufter, P.; Yvon, F.; and Schütze, H. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of EMNLP*, 1627–1643.

Liang, C.; Yu, Y.; Jiang, H.; Er, S.; Wang, R.; Zhao, T.; and Zhang, C. 2020. BOND: Bert-Assisted Open-Domain Named Entity Recognition with Distant Supervision. In *ACM SIGKDD*.

Mayhew, S.; Tsai, C.-T.; and Roth, D. 2017. Cheap Translation for Cross-Lingual Named Entity Recognition. In *EMNLP*, 2536–2545.

Ramesh, G.; Doddapaneni, S.; Bheemaraj, A.; Jobanputra, M.; AK, R.; Sharma, A.; Sahoo, S.; Diddee, H.; J, M.; Kakwani, D.; Kumar, N.; Pradeep, A.; Deepak, K.; Raghavan, V.; Kunchukuttan, A.; Kumar, P.; and Khapra, M. S. 2021. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. arXiv:2104.05596.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

Wu, Q.; Lin, Z.; Karlsson, B.; Lou, J.-G.; and Huang, B. 2020. Single-/Multi-Source Cross-Lingual NER via Teacher-Student Learning on Unlabeled Data in Target Language. In *ACL*.

Wu, S.; and Dredze, M. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *EMNLP-IJCNLP*, 833–844.

Xie, J.; Yang, Z.; Neubig, G.; Smith, N. A.; and Carbonell, J. 2018. Neural Cross-Lingual Named Entity Recognition with Minimal Resources. In *EMNLP*, 369–379.