CL-NERIL: A Cross-Lingual Model for NER in Indian Languages

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

[1] Department of IT, National Institute of Technology Karnataka, Surathkal; [2] Department of CSE, Indian Institute of Technology Guwahati

Overview

- We introduce CL-NERIL, an end-to-end cross-lingual model for NER task on Indian languages in low-resource settings by exploiting parallel corpora of English and Indian languages and an English NER dataset.
- It employs a variant of the Teacher-Student model trained jointly on the pseudo labels of the Teacher model and predictions on weakly-labeled data generated by an annotation projection method.
- Empirical evaluations on manually annotated test sets of three Indian languages: *Hindi*, *Bengali*, and *Gujarati* show over 10% performance improvement compared to zero-shot transfer.

Motivation

- The performance of Named Entity Recognition (NER) models often depends on the amount and quality of annotated data available in a language.
- Indian languages lack such accurate entity annotated corpora to build and benchmark NER systems.

Introduction

- 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 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.
- We use multilingual embedding-based word alignment method to propose an annotation projection method to generate weakly labeled data in the target language.
- This annotation projection method is integrated into an end-to-end framework CL-NERIL along with joint optimization for NER task.

Weakly-Labeled Data Generation

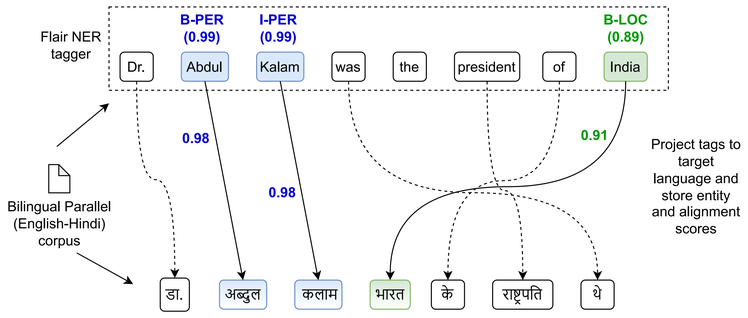

Figure: Weakly-labeled Data Creation

- We determine word and subword alignments between using XLM-R and mBERT embeddings.
- We try to further align the missed entity words (if any) by using the Match algorithm.
- Finally, top 40% sentences based on score constitutes the weakly labeled data.

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

Teacher-Student Model

1. 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.
2. 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.
3. Student model is trained by optimizing MSE loss between the two predicted distributions (considering the Teacher model's predicted labels as soft), and NLL of the generated weak label and the predicted labels (y^{stud}).

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

Experimental Results

Model	bn	gu	hi
Zero-shot [1]	50.67	56.59	70.28
BOND [2]	64.19	79.28	**80.14**
Teacher-Student [3]	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: Benchmarking CL-NERIL against State-of-the-art.

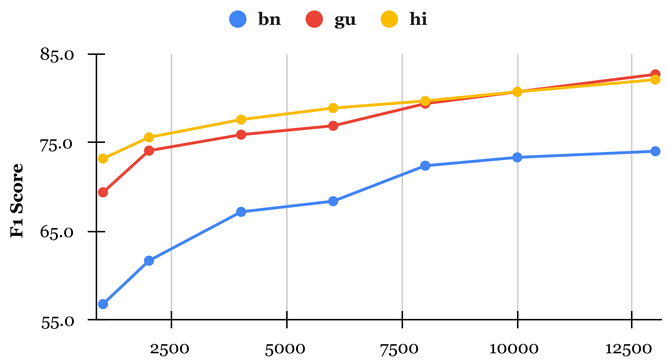

Figure: Size of Weakly-labeled Data

Additional Details

Arxiv Paper

Code

References

[1] S. Wu and M. Dredze, "Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT," in *EMNLP-IJCNLP*, pp. 833–844, Nov. 2019.

[2] C. Liang, Y. Yu, H. Jiang, S. Er, R. Wang, T. Zhao, and C. Zhang, "Bond: Bert-assisted open-domain named entity recognition with distant supervision," in *ACM SIGKDD*, 2020.

[3] Q. Wu, Z. Lin, B. Karlsson, J.-G. Lou, and B. Huang, "Single-/multi-source cross-lingual ner via teacher-student learning on unlabeled data in target language," in *ACL*, July 2020.