BETTER SEQUENCE LABELING FOR LOW RESOURCE LANGUAGES USING ATTENTION AND TRANSFER LEARNING

A thesis submitted in partial fulfillment of the

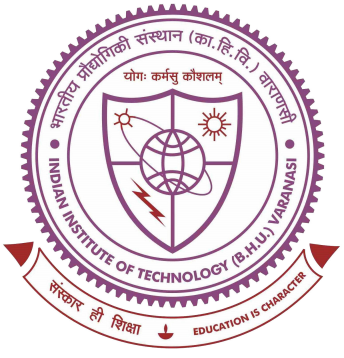
requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

by

RAJESH KUMAR MUNDOTIYA

ROLL NO.: 16071001



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY

(BANARAS HINDU UNIVERSITY),

VARANASI-221 005

February 2022

Chapter 6

Conclusions and Future Scopes

6.1 Proposed Methods and Findings

In this thesis, we have investigated how knowledge from different resources, along with the least amount of annotated data of low resource languages, can be utilized to achieve state-of-the-art results by extending the existing deep learning approaches. Specifically, this thesis has focused on four aspects of sequence labeling for low resource languages: proposing novel variations of deep learning models, integrating handcrafted features, performing transfer learning and providing baseline systems using benchmark seed datasets. Chapters in this thesis are based on these aspects, corresponding to the research questions formulated in the beginning of the thesis. The mapping of these research questions with corresponding chapters is shown in Table 6.1.

Since sequence labeling tasks such as POS tagging, Chunking, NER are the preliminary tasks for processing any new languages in NLP. One of the ways to prevent the language become extinct is to create a baseline tool for such preliminary tasks with seed benchmarked datasets. Chapter 3 provided the baseline tools for Bhojpuri, Maithili and Magahi languages and empirically analyzed the obtained results. Building an automatic tool for these tasks (except Chunking for Magahi) is the first attempt towards these three languages. Recurrent neural networks assume an identical magnitude of sentence words to capture sequential information. Self-attention and monotonic chunk-wise attention helped

TABLE 6.1: Mapping of chapters with their associated research questions

Question	**Chapter**
RQ.1 *How can we reduce the chances of a low resource language becoming an extinct language?*	Chapter 3
RQ.2 *How can we improve the performance of part of speech tagging for low-resource languages by leveraging contextual information and existing auxiliary annotated resources?*	Chapter 3 Chapter 4
RQ.3 *How can we improve the performance of a deep learning architecture for low resource domains by using annotated resources of other domains (not only high but also low resource domains)?*	Chapter 4
RQ.4 *How can we optimize a deep learning architecture's parameters for low resource language by utilizing annotated resources of other languages?*	Chapter 5

to leverage the contextual information is demonstrated in the proposed models, Self Attention based Hierarchical Bi-LSTM CRF (SAHBiLC) and Monotonic Chunk-wise attention with CNN-GRU-Softmax (MCCGS) in chapters 3 (section 3.4.3) and 4 (section 4.6). Here, we focus on incorporating the attention mechanism into the sequence labeling model, as both models outperform the existing baseline approaches. The character-level finetuning of the SAHBiLC model from high-resource language (Hindi) to low-resource language indicates that finetuning is helpful in less training data and complex morphology. However, the knowledge transfer between Hindi to Bhojpuri countered negative transfer even both are similar languages. MCCGS model has been tested on the dataset of multiple languages available on Universal Dependency treebank and Hindi-Urdu treebank for POS tagging.

When endeavoring to integrate linguistics features into the deep learning model, an evident concern arises as to where to add those features as the model comprises numerous layers and each layer is accountable for capturing information. We demonstrated that integrating existing linguistic knowledge of these datasets such as Tense, Case, Gender, Number, Person and Lemma at the label disambiguation improves the performance of the MCCGS model for low resource languages in chapter 4 (section 4.6).

Contrastive training has captured the essential features during model training by minimizing the in-equivalent distribution of data for domain adaptation. In chapter 4 (section 4.7), the proposed model MCCGS extended with contrastive training for POS tagging, which

helps perform domain adaptation with finetuning in single and multi-source settings. The experimental results show that multi-source domain adaptation (irrespective of the data size) yields better results than single-source experimented on Hindi treebank, which comprises Article, Tourism, Conversation and Disease domain and PTB and ARK, TweeBank dataset which comprise Newswire and Tweet domain, respectively.

Few shot learning is an extreme case for low resource settings where few annotated data is assumed for training. We have investigated meta learning algorithms, Model-Agnostic Meta-Learning and Reptile in the context of M-way N-shot learning on POS tagging with the assumption of multilingual scenario in chapter 5. The experimental results have shown that the multilingual meta model trained for few shot POS tagging on finetuning with the low resource language outperforms the NCRF++ framework based baseline. Our results indicate that the pretrained multilingual meta model has captured task-specific features and cross-lingual features that the baseline model is unable to do so due to lack of data. We empirically showed that using multiple scripts for training can improve performance for some languages. We have also shown that having a high resource language from the same script as that of the low resource language during the meta pretraining step helps in improving performance.

6.2 Future Scopes

Despite the presented approaches achieving comparable performance in contrast to earlier work in the corresponding low-resource sequence labeling NLP tasks, there is numerous possible prospect for further research. The limitations of the proposed approaches can be alleviated in the future research direction. The pretrained multilingual or cross-lingual contextual embeddings such as ELMO, BERT, GPT and XLM can be used to handle the negative transfer (as discussed in the section 3.4.4 of chapter 3) and inequivalent distribution of data for domain adaptation (as discussed in the section 4.5 of chapter 4). The domain specific version of these models is available, which can also be leveraged for the in-domain model training. However, building contextual embedding models for

low resource languages such as Bhojpuri, Maithili and Magahi is challenging due to the requirement of a huge monolingual corpus.

The contextual information was encoded with LSTM, GRU and attention mechanism, which showed its significance (as discussed in section 4.5 and section 4.6.3 of chapter 3 and chapter 4, respectively). The fusion of self-attention and monotonic chunk-wise attention can be used to capture better contextual information for these tasks. Transforming existing linguistic features into a one-hot vector to integrate into the model is not a good way to represent features. A real vector must represent those features since deep learning models are sensitive to initialization. Incorporating these features into the word embedding is another direction for improving NLP tools of agglutinative low resource languages. Data augmentation and dealing with label imbalance is another potential future direction of the proposed work.

Contrastive learning minimizes the inequivalent distribution. However, the proposed model in the section 4.5 of chapter 4 considers some common tags as different tags during domain adaptation, which suppresses the model performance for high resource domains. This problem can be overcome by a mapping function that could be a tunable mapping function such as adversarial-discriminator and multi-task or a hard mapping function by exerting linguistic information, considered as future work. Adversarial training allows making the model robust against unknown or OOV words. The adherence to the advantages of contrastive learning and adversarial training can improve the sequence labeling performance. Multi-task learning also improves performance since it requires the aligned annotated corpus of these tasks. So sharing the information among tasks for low resource languages with contrastive learning provides a significant dimension to future work.

The meta learning-based sequence labeling (as discussed in the section 5.6 of chapter 5) has room for further improvement during the model training in multilingual settings; the language has a large amount of annotated dataset dominant over the small amount of annotated dataset languages. The introduction of a balancing factor allows the leverage of information from multiple languages irrespective of the data size.

Working for low resource languages to build a real-world application based on NLP models is still challenging. The thesis contributes a step in this direction and there is a lot to do. The research work of this thesis inspires other researchers to move ahead for the open NLP problems for low resource languages.