Preface

Since the popular spread of the Internet, natural language content in the form of suggestions, opinions, news, and tweets, apart from formal documents, has proliferated at an increasing rate. This has helped in the development of natural language processing (NLP) tools for diverse languages in a variety of domains. Language analysis and understanding is an application of NLP, where machines interpret linguistic information and many other facets of language as humans do. Text classification, syntactic, semantic analysis and entity recognition or linking are examples of problems where errorless decisions are predicted based on language understanding.

Sequence labeling plays a vital role in solving numerous NLP applications such as Machine Translation and Information Extraction etc. In sequence labeling, a category or label is assigned to each member of an instance, where instance denotes a sequence, for example, part of speech (POS) tagging, chunking and named entity recognition (NER). POS tagging assigns a sequence of grammatical categories to the given sentence and Chunking groups them into 'chunks' or what can be called minimal phrases. NER is another sequence labeling problem, which marks proper nouns and other named entities such as Location, Person, Organization, Disease etc. Existing studies have shown that convolutional neural network, recurrent neural network and its variants and conditional random field are the key components for deep learning models. They, along with pre-trained embedding, provide good results for these sequence labeling problems. However, these solutions do not generalize for low resource languages, which are characterized by very little benchmarking seed amount of (un)-annotated data. Transfer learning and multilingual learning do, however, enable gains in performance. Improving the performance of sequence labeling models using existing knowledge is one of the prominent research areas in NLP for low resource languages. This thesis focuses on four aspects of this: proposing deep learning architectures, integrating handcrafted features, performing transfer learning, and providing baseline systems using benchmark seed datasets for some low resource languages.

Bhojpuri, Maithili and Magahi are low resource languages and widely spoken in central north-eastern India, belonging to the Indo-Aryan language family. Creating an annotated corpus for POS tagging and chunking and then building an initial automatic tool for these languages is the first attempt towards the development of language technology tools for these languages. The annotated corpus used to develop POS taggers and chunkers, based on various traditional machine learning algorithms (trigrams 'n' tags, conditional random fields (CRF), maximum entropy Markov model and structured support vector machine) and more recent state-of-the-art deep learning model which comprise long short-term memory (LSTM), convolution neural network (CNN) and CRF, named as LSTM-CNN-CRF model, has been used.

A robust deep learning model has been proposed that leverages character level information to deal with out-of-vocabulary and an attention mechanism to capture the inter-dependence of input words known as Self-Attention-based Hierarchical Bi-LSTM CRF (SAHBiLC) model. Model transfer is one of the ways of transfer learning by which the learned knowledge (parameters of deep learning model) can transfer from one model to another model; hence initially, the model has been trained on Hindi and then transferred to Bhojpuri, Maithili and Magahi: this setting being referred to as the fine-tuned version of SAHBiLC model, named as Fine-SAHBiLC. SAHBiLC and Fine-SAHBiLC outperform previous results for Bhojpuri and Maithili, Magahi, respectively, for both tasks. This shows that finetuning is effective for such languages which have complex morphology and lesser training data.

Despite the success of the proposed model, we compare the results of CRF and LSTM-CNN-CRF for the NER datasets of these languages since the number of named entities are relatively few in the data, that is, the datasets of these languages are highly imbalanced. Here, the deep learning baseline provides a better result for Magahi only due to fewer intermediate entities (tags). We observe that the obtained results are consistent with the number of named entities in the datasets rather than with the total size of the dataset in the number of tokens.

Traditional machine learning algorithms require feature engineering. Linguistic information such as affixes, lemma, adjacent words is commonly used for these sequence labeling problems. Recent deep learning models combine the forward and backward word information captured by the recurrent neural network or its variants for POS tagging. However, it assumes an identical magnitude of words of a sentence to capture sequential information. The information of contextual words to the current word plays a vital role in capturing the

non-continuous relationship. Based on the same assumption, we have proposed a Monotonic chunk-wise attention model with CNN-GRU-Softmax (MCCGS), a deep learning model that captures this essential information. This architecture consists of Input Encoder (IE), which encodes word and character-level information, Contextual Encoder (CE), which assigns weightage to adjacent word and Disambiguator (D), which resolves intra-label dependencies as core components. Later on, this proposed model gets integrated with morphological information as an external feature. Tense, Case, Gender, Number, Person and Lemma considered external features integrated into the core components of MCCGS architecture as MCCGS-IE, MCCGS-CE and MCCGS-D. The MCCGS architecture and its variants are validated on datasets for 21 languages from the universal dependency (UD) treebank. Compared to the state-of-the-art position-aware self-attention-based model, the MCCGS model has improved by 0.29% in mean accuracy. The MCCGS model improved further by 2.98% due to unification of features at the Disambiguator component (MCCGS-D).

If we assume that the distribution of source and target data for a language is inequivalent, then modeling over these settings is widely known as domain adaptation. The contrastive training minimizes the difference of inequivalent distribution. Contrastive training has been tried as a robust approach that captures the essential features during model training, and based on this, contrastive-MCCGS model architecture has been proposed for POS tagging. It learns optimal features in a low resource regime.

We experimented on the datasets of four domains, Article, Conversation, Disease and Tourism, of the Hindi treebank, Tweet domain from TweeBank, Newswire domain from Penn TreeBank (PTB) and Tweet domain from ARK and compared it with several state-of-the-art models. The CMCCGS model has been further extended to domain adaptation by using single and multi-source domain adaptation to allow fine-tuning. Compared to single-source adaptation, significant improvements have been observed after using multi-source domain adaptation. Since it is multi-source domain adaptation, for experiments based on the deep learning models, the effect of layer freezing was observed. Multi-source domain adaptation provides competitive results over the baselines. Very low resource domains such as Tourism, Disease and tweet domain of TweeBank and ARK have shown improvement in accuracy by Article, Article and Tourism (multi-source), PTB, and PTB and TweeBank (multi-source) as source domain, respectively. However, the Conversation domain has a negative impact on domain adaptation.

Devising deep learning networks that perform well on low resource languages is challenging

since very little labelled data or human annotated data is available for these languages. One of the possible solutions to overcome this challenge is to use multilingual training with accurate parameter transfer. We present a multilingual meta learning based algorithm to learn representations for POS tagging of low resource languages. Meta learning algorithms learn a model from the distribution of tasks that can only adapt to a previously unseen task with a few examples. Two meta-learning algorithms, model agnostic meta learning (MAML) and Reptile, are used to train the model. We conduct extensive experiments on POS tagging using nine languages from two different scripts, out of which the model is trained on eight languages via meta learning and fine-tuned on an unseen low resource language. Data for all nine languages is obtained from the Universal Dependencies Treebank. Empirical results have shown that a model trained via multilingual meta learning can learn cross-lingual features, cross-script features and task-specific features that can easily be transferred into previously unseen low resource languages. Our Approach gains significant improvement over a robust baseline on a low resource language. An experiment for few shot POS tagging verifies the learned representations.