# Abstract

In recent years, Natural Language Processing (NLP) has made significant strides, leading to the development of advanced models and techniques capable of understanding and generating human-like language. However, the benefits of these advancements have primarily favored English, leaving many other languages at a disadvantage. Low-resource languages, such as Indic languages, present unique challenges in NLP due to limited linguistic data and resources. Collecting and annotating extensive data for training in these languages is daunting and economically burdensome. This scarcity poses a significant barrier to developing robust and accurate models, as contemporary NLP techniques heavily rely on large datasets for effective training.

Multilingual encoder-based models like mBERT and XLM-R have shown promise in various low-resource tasks. These pre-trained models possess distinctive features advantageous for addressing challenges in low-resource languages. Their multilingual capabilities effectively handle language diversity, while shared representations empower them to learn and extract universal linguistic features across multiple languages.

Transfer learning approaches are crucial in addressing challenges associated with

low-resource languages in NLP. These strategies involve pre-training models on languages with abundant linguistic resources and fine-tuning them on the target low-resource language. Transfer learning is essential for low-resource languages, as it enables the utilization of pre-trained models, overcomes data scarcity, handles out-of-vocabulary words, and facilitates domain adaptation tasks. Meta-learning, or learning to learn, is a prominent transfer learning paradigm. The Model-Agnostic Meta-Learning (MAML) algorithm, a well-known meta-learning technique, trains models to identify parameter initializations that enable rapid adaptation to new tasks with minimal updates. Additionally, knowledge distillation proves to be a practical transfer learning approach, particularly beneficial in scenarios with a shortage of labeled data. It involves transferring knowledge from a teacher model trained on a larger dataset to a smaller student model, facilitating effective training with reduced reliance on labeled data. Cross-lingual training is another valuable strategy in transfer learning, especially for low-resource languages. This approach leverages knowledge from high-resource languages and applies it to tasks in low-resource languages, capitalizing on linguistic similarities, facilitating knowledge transfer, and enabling the transfer of multilingual competence.

This thesis explores various transfer-learning methodologies, employing diverse pre-trained models with multilingual encoders, to address key challenges in natural language processing for low-resource languages across three critical tasks: Event Detection (ED), Question Generation (QG), and Neural Machine Translation (NMT).

In Event Detection (ED), the scarcity of annotated data poses a significant obstacle. Cross-lingual ED aims to mitigate this challenge by transferring knowledge across languages to enhance overall performance. This thesis introduces a model-agnostic meta-learning approach for few-shot cross-lingual ED, enabling rapid adaptation to new low-resource languages. Evaluation of four Indian languages demonstrates significant performance improvements over the base model. Question Generation (QG) presents another crucial challenge in the Natural Language Generation (NLG) domain, particularly in low-resource settings. Leveraging Multilingual BERT (mBERT) for few-shot QG with cross-lingual transfer remains an open question. This research explores the performance of mBERT in few-shot QG and investigates the efficacy of applying meta-learning to enhance results. The proposed approach consistently outperforms the base model in few-shot settings, as demonstrated in Bengali and Telugu languages using the TyDi QA dataset. Human evaluation further confirms the effectiveness of the approach. Neural Machine Translation (NMT) remains a formidable challenge for low-resource languages, exacerbated by the limited language support of pre-trained models like mBART-50. This thesis proposes a framework that leverages pre-trained language models and knowledge distillation in a seq2seq architecture to address this challenge. The framework, evaluated on six low-resource Indic language pairs, achieves significant BLEU-4 and chrF improvements over baselines, with human evaluation confirming its effectiveness.

Overall, this thesis contributes to a deeper understanding of how transfer-learning techniques, meta-learning, and knowledge distillation can effectively improve NLP tasks for low-resource languages, paving the way for more robust and efficient multilingual NLP systems.