Chapter 1

Introduction

Language is a medium of communication and it is often considered as a defining characteristic of human intelligence. In the past few years, the availability of natural language text on the Internet has increased very fast due to, for example, the sharing of views by people in their languages. These views are expressed in a variety of languages on the Internet. Such languages might be for which vast quantities of digital resources are available and they can be called resource rich languages. English, Japanese and Chinese are a few of such languages, and the availability of text in them is getting more and more richer in the form of opinions, suggestions, articles, blogs and social media content, apart from conventional kinds of texts such as those which are publicly printed. Most of the content on the Internet belongs to such resource rich languages. The content of resource poor languages has also increased, but not that much.

The rapid increase in language content on the Internet has also come in the form of a platform for resource poor languages. However, in the last few years, the domain of some languages has also been shrinking mainly due to the pressures of globalization. As a result, many languages are becoming extinct, and some are on the verge of becoming endangered languages. According to Ethnologue, 2,926 languages had become endangered by the end of 2020[1] in the world. This is related to the fact that language technology has not yet been established for most languages, even if many speakers are still alive of languages on

[1] https://www.ethnologue.com/guides/how-many-languages-endangered

the brink of extinction. In the newest study of Ethnologue 2021, there are approximately 7,139 living languages[2] in the world. Out of which 88% of the world's population use the top 200 languages as a native or second language[3]. Language technology development for some of these languages (about 23) has received disproportionate attention in terms of the number of languages. According to Crystal [48], at the end of this century, half of the world's languages would be extinct if this trend continues. Language is a key part of the identity of an individual and their society, and along with culture and heritage it is like a glue that binds one society internally or to another society or the global society. For these reasons, documentation, conservation and technology development of languages is becoming a significant priority these days.

Defence Advanced Research Projects Agency (DARPA) Low Resource Languages for Emergent Incidents (LORELEI)[4], DARPA TIDES (Translingual Information Detection, Extraction, and Summarization), IARPA Babel[5], Research on English and Foreign Language Exploitation (REFLEX) LCTL program, National Institute of Standards and Technologies (NIST) Language REcognition (LRE)[6], the Digital Language Diversity Project (DLDP)[7] and Enabling Languages with Low Resources (ELLORA)[8] are such programs which have been initiated with the same motivations. According to these programs, languages have been broadly categorized into different categories based on the speakers (native or second language) and the availability of digitized resources.

Thus, if numerous speakers and plenty of digitized resources are available worldwide for a language, such a language is considered a high resource languages, for example English. Researchers have been using digitized resources with the help of native speakers to develop automation tools to achieve advancement in language technology. However, sometimes these resources do not become available in ample amounts despite the accessibility of native speakers. Such languages are considered low resource languages. For example, Bengali and Indonesian come under the worldwide top 10 spoken languages, but resources

[2] https://www.ethnologue.com/ethnoblog/gary-simons/welcome-24th-edition
[3] https://www.ethnologue.com/guides/ethnologue200
[4] https://www.darpa.mil/program/low-resource-languages-for-emergent-incidents
[5] https://www.iarpa.gov/index.php/research-programs/babel
[6] https://www.nist.gov/itl/iad/mig/language-recognition
[7] http://www.dldp.eu/en/content/project
[8] https://www.microsoft.com/en-us/research/project/ellora/

have not been adequately available for technology development for them so far. In addition to these categories, endangered languages are defined as languages that have disappeared and are disappearing due to a lack of native speakers.

These digital language resources are processed and understood by the machine, including their contexts, to create artificial intelligence, computational linguistics and natural language processing. Natural language processing facilitates machines to process and understand human language. Proposed by Alan Turing in the 1950s, the Turing test was for long considered the criterion for intelligence for machines, i.e., machines which can interpret and generate natural language text or speech. From a research point of view, the applications of natural language processing when a human is always in a loop can be generalized in the following ways:

- **Human-Machine Communication**: Adequate knowledge about human language and performance during machine-based interaction. Speech recognition or synthesis, Dialogue generation and Question answering are examples of human-machine communication.
- **Human-Human Communication**: The application used to connect humans without interpreting the human language, viz. Machine translation.
- **Analyze and Understand Language**: The linguistic information and many facets of language are interpreted by machines as humans do. Text classification, Syntactic analysis, Semantic analysis and Entity recognition or linking are examples of problems where errorless decisions are entirely anticipated based on language understanding.

The most earlier approaches for natural language processing (NLP) applications were symbolic approaches that use rules and lookup dictionaries [16]. Creation of these rules and dictionaries by a human had been a laborious task for the symbolic approach. This approach is hard to generalize with unknown inputs because they are language, domain and task-specific [248], even if languages are generative in nature according to Chomsky [43].

The statistical approach facilitates generalizing to a certain extent on unknown inputs. In this approach, the rules are not mentioned explicitly. The machine learns those rules with the help of mathematical modelling on the provided inputs and outputs [145].This approach also employs feature engineering to achieve better results, which requires considerable human input.

A revolutionary growth of machine capabilities, processing power and data has been recorded since 2010. A neural-based approach provides state-of-the-art results for various NLP applications such as Machine translation, Text classification and Sentiment analysis by implicit learning of multi-layered features that overcome feature engineering requirements. Consequently, human efforts have focused on ascertaining the most appropriate architecture and training settings for a single or joint application. The neural-based approach is also known as the Deep learning approach. Deep learning requires extensive annotated data for getting state-of-the-art results.

1.1 Research Objectives

Extending the existing deep learning approaches to low resource languages is a thrust area of NLP. The knowledge from different resources along with the least amount of annotated data of these languages is utilized to achieve state-of-the-art results, which is a fruitful research direction. This thesis studied following four research questions that emphasise on the processing of low resource language and obtaining an accurate deep learning architecture:

1. *How can we reduce the chances of a low resource language from becoming an extinct language?* Language analysis and understanding based applications are preliminary applications to work with any new languages, which also help to link with similar high resource languages. On the same assumption, low resource languages can be transformed into resource-rich languages with the help of these applications.

2. *How can we improve the performance of NLP task's for low-resource languages by leveraging contextual information and existing auxiliary annotated resources?* This

question aims to make a model more robust by utilizing handcrafted features on which machine learning models work.

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)?* This question aims to study how the combined use of language resources from diverse sources (languages or domains) enhances the performance of deep learning models through a transfer learning technique.

4. *How can we optimize a deep learning architecture's parameters for low resource language by utilizing annotated resources of other languages?* This question aims to re-utilize the trained parameters of other languages simultaneously to generalize the model with few annotated resources.

1.2 Scope

NLP has a variety of tasks that have to be successfully completed for various linguistic problems. Most of them are related to syntax and semantic analysis. For example Text summarization, Part of speech tagging, Chunking, Named entity recognition, Parsing, Coreference resolution, Machine translation, Sentiment analysis, Optical character recognition, Natural language understanding, Natural language generation, Morphological analysis, Morphological generation, Word segmentation, Word sense disambiguation, Speech recognition, Speech segmentation, Speech synthesis, and Question answering are some such tasks.

The thesis's scope is confined at language analysis and understanding based applications, in particular, syntactic analysis or information extraction for low resource languages. The tasks dealt with in this thesis are as follows:

- Part-of-speech Tagging: Part-of-speech (POS) tagging analyzes the syntactic structure of the text and assign grammatical categories to the words in a sentence, such as Noun, Verb, Adverb and Adjective.

- Chunking: Chunking links POS tagged words into groups of words or 'chunks', which can be roughly defined as minimal phrases or minimal constituents such as Noun Phrase, Verb Phrase and Adjectival Phrase.

- Named entity recognition: Named entity recognition (NER) is one of the preliminary task, which marks proper nouns and other named entities such as Location, Person, Organization, Disease etc.

These tasks can be intermediate steps of an interlingua based machine translation system, which exploits source and target language analysis. They can also independently useful for other applications, either for providing features or for linguistic analysis for researchers. The reason for choosing these tasks is because they appear in most NLP pipelines, and an advancement in NLP for low resource languages cannot be realized without recourse to these tasks.

1.3 Contributions

This thesis focuses on four aspects: proposing deep learning architectures, integrating handcrafted features, performing transfer learning and provide baseline systems by using benchmark seed datasets to low resource languages for sequence labeling problems.

One of the primary goals of NLP is to transfigure low resource languages to resource-rich languages. One of the prevalent methods is to build seed annotated datasets for such languages so researchers can utilize them to develop NLP tools. In this respect, We provide baseline tools for POS tagging, Chunking and NER to three low resource languages, Bhojpuri, Maithili and Magahi as well as leverage the transferred knowledge from a similar language, Hindi and analyze the effect of negative transfer.

We analyze the improvement in performance after incorporating an attention mechanism in deep learning models since it allows us to consider the contextual information for sequence labeling problems for low resource languages. We demonstrate that the integration of existing hand-crafted features into the proposed deep learning architectures improves the

performance and evaluates the best integration layer. These features allow curbing the wrong prediction of target labels for out of vocabulary words, which is a prevalent issue in low resource languages due to limited types in an annotated dataset. Contrastive learning is another approach to generate synthetic vectors during model training which again helps to improve the performance.

The proposed architecture with contrastive learning is applied to domain adaptation by finetuning from one domain to another, where the source domain could be single or many. Later on, the same idea has extended to multiple languages where meta-learning techniques have been used to optimize the parameters based on multiple languages towards a target language. Here, multiple languages can be a group of low resource languages which use the same or different scripts. We demonstrate that multiple languages improve the sequence labeling performance for low resource languages, irrespective of their annotated data size and scripts.

1.4 Chapterization

The thesis chapters are as follows:

- Chapter 2 describes preliminaries such as machine learning models, deep learning models, transfer learning techniques, integration of handcrafted features into deep learning models and a brief overview of the existing work, based on low resource settings for sequence labeling problems. It also sheds light on supervised or minimally supervised work on an associated tasks for Indian languages.
- Chapter 3 describes machine learning based baseline experiments and reports evaluation over sequence labeling problems such as POS tagging, Chunking and NER for three Indian low resource languages (Bhojpuri, Maithili and Magahi). Proposes a novel architecture based on the attention mechanism to leverage sentence level information for these languages.
- Chapter 4 explores the proposed architecture for integrating contextual information in the deep learning model through an attention mechanism for low resource POS

tagging. Later on, the proposed architecture has been extended in two ways, where the first way comprises the handcrafted features and the second way includes contrastive learning for the POS tagging. The first way aims to find the best layer to integrate handcrafted features in the proposed deep learning model. Exploring the single and multi-source domain adaptation on the proposed contrastive-based deep learning model is the aim of the second way.

- Chapter 5 provides the details of proposed architectures based on meta learning techniques to low resource POS tagging. The meta learning techniques are used to design multilingual learning for low resource languages. Based on the availability of limited annotated data, the meta learning model is trained for few shot POS tagging and is finetuned on a low resource language's dataset. It is empirically observed that the learned meta model can capture task specific, cross-lingual and cross-script features.

- Chapter 6 concludes the thesis by summarizing findings, analysis of research outcomes, and future directions.