ANALYSIS OF PART OF SPEECH TAGGING FOR MIZO LANGUAGE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

MORREL V.L. NUNGSANGA

MZU REGN NO : 1807566

Ph.D. REGN NO : MZU/Ph.D/1298 OF 03.07.2018

DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF ENGINEERING AND TECHNOLOGY
SEPTEMBER, 2023

ANALYSIS OF PART OF SPEECH TAGGING FOR MIZO LANGUAGE

BY

MORREL V.L. NUNGSANGA
DEPARTMENT OF INFORMATION TECHNOLOGY

Supervisor : Prof. L. Lolit Kumar Singh
Joint-Supervisor : Dr. Partha Pakray

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Information Technology of Mizoram University, Aizawl.

CERTIFICATE

This is to certify that the thesis entitled "*Analysis of Part of Speech Tagging for Mizo Language*" submitted to the Mizoram University in the Department of Information Technology under the School of Engineering and Technology, in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Information Technology is a record of research work carried out by Mr. Morrel V.L. Nunsanga under my personal supervision and guidance.
All helps received by him from various sources have been duly acknowledged. No part of this thesis has been reproduced elsewhere for award of any other degree.

Dr. L. Lolit Kumar Singh
Supervisor & Professor
Dept. of Electronics & Communication Engineering
Mizoram University,
Aizawl-796 004

Place:

Date:

Dr. Partha Pakray
Joint Supervisor & Assistant Professor
Dept. of Computer Science & Engineering
National Institute of Technology Silchar,
Assam-788 010

Place:

Date:

DECLARATION

Mizoram University

September, 2023

I, Morrel V.L. Nungsanga, hereby declare that the subject matter of this thesis is the record of work done by me, that the contents of this thesis did not form basis of the award of any previous degree to me or to do the best of my knowledge to anybody else, and that the thesis has not been submitted by me for any research degree in any other University/Institute.

This is being submitted to Mizoram University for the degree of Doctor of Philosophy in Information Technology.

(MORREL V.L. NUNSANGA)

(Prof. L. LOLIT KUMAR SINGH)

Supervisor

(Dr. R. CHAWNGSANGPUII)

Head

(Dr. PARTHA PAKRAY)

Joint Supervisor

ACKNOWLEDGEMENTS

Foremost, heartfelt appreciation goes to the supervisor, Prof L. Lolit Kumar Singh, the Department of ECE, Mizoram University. The consistent guidance, profound insights, and continuous encouragement provided by Prof. Lolit have been invaluable in shaping the direction of the research.

I would also like to extend my sincere appreciation to my joint supervisor, Dr Partha Pakray, Department of CSE, NIT, Silchar, for his insightful contributions and support. His extensive knowledge in the area of Natural Language Processing and his willingness to engage in meaningful discussions have significantly enriched the quality of this research.

I'm grateful to the Department of Information Technology, Mizoram University, for fostering an ideal research environment and providing crucial resources.

I acknowledge the Department of ECE, MZU, for my initial admission, and the subsequent opening of the Ph.D. program in the Department of IT, which allowed my research journey to continue seamlessly.

My heartfelt thanks to all professors and colleagues for their expertise and feedback, shaping my research at various stages.

I also would like to express my heartfelt thanks to my family and friends for their unwavering support and understanding throughout this journey.

Last but not least, my heartfelt gratitude goes to the All Mighty God, whose divine presence guided me through uncertainties and granted me resilience.

(MORREL V.L. NUNSANGA)

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

The advancement of technology continues to play a transformative role in human life, and as a consequence, the significance of language processing has reached unprecedented heights. In today's world, the ever-expanding wealth of information and knowledge available in numerous languages, coupled with the escalating cross-cultural interactions, has amplified the need for language technology like never before.

Charles Babbage's invention of the computer was considered a technological revolutionary milestone in human history [1]. Subsequently, significant technological advancements have occurred, enabling computers to handle much more complex tasks. Initially, computer instructions and data were conveyed using only binary code, 1s and 0s, making the process highly intricate and time-consuming. However, the demand for more user-friendly and efficient computer interactions in our fast-developing world prompted the exploration of natural language processing (NLP) as a sought-after technological and linguistic breakthrough.

The foundation for this transformative technology can be traced back to the work of Swiss linguistics professor Ferdinand de Saussure, who taught classes at the University of Geneva between 1906 and 1911 [2]. Saussure's revolutionary idea conceptualized languages as "systems" that facilitate communication, binding societies together through shared language and common social norms. The profound impact of his teachings led two of his former students, C. Bally and A. Sechehaye, to publish his most significant work, "Course in General Linguistics" (Cours de linguistique générale), posthumously in 1916 [3]. This seminal publication laid the groundwork for modern linguistics and played a fundamental role in shaping the principles of natural language processing.

This chapter provides an informative overview of Natural Language Processing (NLP), specifically focusing on Part-of-Speech (POS) tagging. It explores the fundamental concepts of NLP, delves into the significance and applications of POS tagging, discusses various types of POS tagging techniques, and provides insights into the motivations driving this research. Furthermore, it outlines the specific objectives this research aims to achieve, paving the way for a comprehensive understanding of the challenges and opportunities associated with POS tagging in NLP. By presenting this holistic perspective, this chapter sets the stage for the subsequent sections, laying the groundwork for a thorough exploration of POS tagging for the Mizo language.

1.1 Natural Language Processing

Natural Language Processing (NLP) is a rapidly evolving field that encompasses the processing of natural languages, both written and spoken, to extract meaningful information using computer programs. Situated at the intersection of linguistics, computer science, and artificial intelligence, NLP enables the analysis of large volumes of human language data and facilitates its translation into a format that computers can comprehend [4]. The ultimate goal of NLP is to enable computers to achieve a level of linguistic comprehension comparable to that of humans.

A key area of focus within natural language processing is the development of computational models that mimic human language processing. This pursuit serves two primary purposes. Firstly, it aims to create automated language processing tools that can assist in various tasks, such as information retrieval, sentiment analysis, machine translation, and speech recognition [5]. These tools not only enhance efficiency but also enable the handling of vast amounts of textual data that would otherwise be overwhelming for humans to process manually.

Secondly, the development of computational models of human language pro-

cessing aims to deepen our understanding of human communication [5]. By studying how computers can interpret and generate human language, researchers gain insights into the intricacies of linguistic structures, semantics, and pragmatics. This knowledge contributes to fields like linguistics, cognitive science, and psychology, providing valuable insights into the nature of human language itself.

Unstructured textual data constitutes a significant portion of the information available today. To extract meaningful insights from such data, it is essential to employ techniques of text analysis and transformation through NLP. By converting unstructured text into structured data, NLP enables the application of various analytical methods, including machine learning and data mining, to extract patterns, sentiments, and other valuable information [6]. This structured representation allows for easier integration with other data sources and facilitates advanced analysis and decision-making processes.

The advancement of NLP and its integration into various applications have led to its increasing presence in our daily lives. From virtual assistants and chatbots to language translation services and sentiment analysis in social media, NLP has become an indispensable component of many technological solutions [7]. The rapid progress in this field demands that we harness its promising potential to enhance human-computer interaction, enable better information retrieval, and drive innovation in diverse domains ranging from healthcare and finance to education and entertainment.

1.2 Part of Speech (POS) Tagging

POS tagging, also referred to as part-of-speech tagging, serves as a fundamental component within language processing pipelines. This crucial task simplifies various intricate challenges in the field of computational linguistics. It involves the identification and labeling of each word in a text, assigning it to its corresponding grammatical category like noun, pronoun, adjective, etc., or a lexical class,

based on its context within the sentence and its definition within the corpus [8]. Essentially, POS tagging associates a specific part of speech with each word in a sentence, providing valuable insights into its usage within the sentence or phrase context. By conveying information about the grammatical role of each token, it enables more advanced analysis and manipulation of the text.

A part-of-speech tag, also known as a POS tag, serves as a distinctive label assigned to every word or token found in a text corpus, as detailed by Santos et al. [9]. The complete set of POS tags within a given corpus constitutes a tagset. These tags are designed to indicate the grammatical category to which tokens belong. Tagsets may exhibit variations across different languages, reflecting diversity among unrelated languages and similarities among related ones, albeit with occasional exceptions. These tagsets are carefully constructed to abstractly represent the morphological characteristics of text within the corpus, thus facilitating subsequent linguistic analyses.

When implementing a POS tagger, it takes a sequence of text in a specific language as input and allocates the appropriate part-of-speech tag to each word or token within the sequence. The accuracy of the POS tagger is often considered important, as it can contribute to improving the performance of subsequent phases in various NLP tasks [10].

Managing ambiguity represents a central hurdle in the task of POS tagging. Many words in a language have multiple senses, making it challenging to determine the correct POS tag without considering other types of linguistic information like syntax, semantics, and world knowledge. As such, achieving flawless tagging is difficult, but it's still possible to attain a high degree of accuracy, which proves highly valuable for practical purposes in natural language processing. Researchers and developers continuously strive to improve the accuracy of POS taggers by leveraging advancements in machine learning and linguistic modeling, leading to more refined and reliable language processing systems.

Identifying part-of-speech (POS) tags presents a complex undertaking, pri-

marily due to the inherent ambiguity present in language. Words possess multiple meanings and can assume varying parts of speech contingent upon their contextual usage. To illustrate this point, let's examine two example sentences:

"I saw a bat."

"I cannot bat at number three."

In these sentences, the word 'bat' carries distinct meanings. In the first sentence, it functions as a noun, referring to the flying mammal. However, in the second sentence, 'bat' acts as a verb, indicating the action of hitting a ball in a sports context. This contrast highlights the importance of considering the context in understanding word usage and determining the appropriate POS.

Let's explore the Mizo language and the word *'lei'*, which has different meanings due to tone variations as given below:

Lei (Short mid-tone) – to buy - Verb.

Lei (Short high-tone) -tongue - Noun.

Lei (short low-tone) - bridge - Noun.

In Mizo, the different tones associated with *'lei'* result in different interpretations. Pronounced with a short mid-tone, it signifies the verb 'to buy.' With a short high-tone, it represents the noun 'tongue,' while a short low-tone pronunciation corresponds to the noun 'bridge.' These tone-dependent variations further highlight the complexities of language and the challenges in accurately assigning POS tags.

Similarly, different languages may differ in their degree of uncertainty and complexity, with some exhibiting more ambiguity than others due to their grammatical structures or extensive lexicons with multiple-word meanings. Therefore, computational linguists must continually refine existing models and develop novel techniques to effectively address the intricacies and challenges of ambiguity in language processing. Having precise POS tagging is a viable approach to tackle language ambiguity in processing. It plays a crucial role in computational lin-

guistics knowledge by facilitating efficient language analysis and improving the performance of diverse natural language processing tasks [11]

1.3 Applications of Part-of-Speech Tagging

Applications of part-of-speech (POS) tagging encompass a broad spectrum within the field of computational linguistics. It involves providing grammatical information for each token in a corpus, which can significantly impact various modules in the natural language processing (NLP) pipeline [12]. POS tagging serves as an initial step in several linguistic applications, including text-to-speech synthesis [13] [14] [15], machine translation [16] [17], word sense disambiguation [18] [19] [20], speech recognition [21] [22], information retrieval and extraction [23] [24] [25], question and answering [26], text summarization semantic analysis, etc. The following discussion will delve into some of these applications in more detail.

(i) *Text to speech:* Parts of speech provide good information about the word and its neighbors, which is helpful in the language model for speech recognition. For instance, the word "read" has two distinct pronunciations depending on the tense. In the present tense, the word "read" is pronounced similarly to the word "reed"; however, in the past tense, the word is pronounced similarly to the color "red." If proper tagging is done on the text, it will help to provide correct pronunciation.

(ii) *Word-sense disambiguation (WSD):* It is the task of identifying the correct meaning of a particular word used in a text when it has multiple meanings. Consider the word "address" in the following two sentences :

I address the gathering.

I could not find the address.

"To talk to" is what "address" means in the first statement, which is a verb. In the second sentence, it is a noun that refers to the term "location." A POS

tagger can assist the machine in determining the context in which the word is being used. If this tagging information is available, the exact interpretation can be obtained.

(iii) *Speech recognition:* POS tagging is often used as a pre-processing step in speech recognition systems. It aids in improving the accuracy of speech-to-text conversion by mapping the acoustic signals to appropriate word sequences, considering the grammatical structure.

(iv) *Machine translation:* In machine translation systems, POS tagging helps in disambiguating words with multiple possible translations. By considering the grammatical roles of words, the translation model can generate more accurate and contextually appropriate translations. Numerous experiments [27] [16] [17] [28] have demonstrated the usefulness of POS tagging in providing additional syntactic information for Neural machine translations(NMT) models, which may help to reduce linguistic ambiguity.

(v) *Information retrieval and extraction:* In the retrieval system, it is crucial to determine the intended meaning of each vocabulary word. POS tag is a potentially strong signal for word sense disambiguation. The result of the retrieval system can greatly be refined by adding the POS tag information in the retrieval system. Numerous research teams have made use of the POS tag's benefit in the retrieval system [29] [30] [24] [25] [23]

(vi) *Question & answering:* POS tagging can assist in question-answering systems by identifying the grammatical structure of questions and mapping them to relevant parts of a document or knowledge base. This helps in retrieving and extracting the most appropriate answers

1.4 Types of Part-of-Speech Tagging

At the topmost level, POS tagging techniques can be broadly divided into Supervised tagging and Unsupervised tagging [31]. Unsupervised tagging does not use any tagged corpus and uses advanced algorithms for creating the tag set as well as deriving the transformation rules. Supervised tagging involves the use of a tagged corpus which is used for training the POS tagger. Under Supervised and Unsupervised approaches, different approaches for POS tagging exist.

In fact, different researchers may hold varying viewpoints and methodologies when it comes to the classification and categorization of POS tagging techniques. While there are commonly acknowledged categories like rule-based, probabilistic, neural network-based, or transformation-based tagging, authors may diverge in their specific subcategories and terminology. Some researchers might focus on the algorithms or models employed, while others may emphasize linguistic features or rule-based strategies. The following discussion highlights some commonly accepted classifications:

(i) **Rule-based approach:** The rule-based approach for POS tagging involves manually crafted rules that assign POS tags to words based on their context and neighboring words. These rules are typically designed by linguists and language experts. While this method can be accurate for specific domains or languages, it requires significant manual effort and may not work well in different contexts. One advantage is that it reduces the need for extensive data storage, relying instead on a set of predefined rules. However, creating these rules demands in-depth knowledge of the language, and it can be challenging to encompass all linguistic nuances. Furthermore, since the rules are predefined and based on known words, they may not have the ability to assign POS tags accurately to words that are not present in the rule set. Despite these limitations, the rule-based approach has been utilized in various research papers, including references such as [32], [33], [34], [35], [36]

(ii) **Stochastic approach:** Stochastic approach in POS (Part-of-Speech) tagging refers to a statistical or probabilistic method used to assign grammatical labels (part-of-speech tags) to words in a given sentence or text. In this approach, the POS tag for each word is determined based on the probability of observing a particular tag given the word and its context [37].

Stochastic POS tagging typically involves training a statistical model on a large annotated corpus, where each word is labeled with its correct POS tag. The model learns the statistical patterns and dependencies between words and their corresponding POS tags. It estimates the probability distribution of tags for each word based on its context, such as the surrounding words or the words within a certain window. The models employed in stochastic taggers include Hidden Markov Models (HMM) [38] [39] , Support Vector Machines (SVM) [40] [41], n-gram , Decision trees [42] [43], Maximum Entropy Markov models (MEMM) [44] [45], and Conditional Random Fields (CRF) [46] [47], [48] [49] [50].

Stochastic POS tagging has proven effective in addressing ambiguity and achieving decent accuracy in various natural language processing tasks. However, its performance heavily relies on the quality and size of the training corpus to effectively handle new data. In some cases, the tagger may generate tag sequences that do not conform to grammar rules.

(iii) **Hybrid approach:** The hybrid approach to POS tagging is a methodology that combines the strengths of multiple techniques to achieve more accurate and robust part-of-speech (POS) tagging results. It integrates two or more different POS tagging methods, typically combining statistical and rule-based techniques. In a hybrid approach, the system utilizes the advantages of each individual method to compensate for the weaknesses of the others. By merging these different techniques, the hybrid approach aims to improve the overall accuracy and performance of POS tagging. Numerous POS taggers employing the hybrid approach have been developed, including those mentioned in references [51], [52], [53] [54].

(iv) **Neural Network approach:** A neural network, also known as an artificial neural network (ANN), is an artificial intelligence technique that emulates the human brain's data processing capabilities. It utilizes interconnected nodes or neurons organized in layers and continually enhances its performance by learning from errors [55].

The ANN comprises interconnected nodes or artificial neurons arranged in layers, responsible for processing and transmitting information [55]. These nodes receive input signals, perform computations, and generate output signals. The connections between nodes possess weights that dictate the strength of influence between them [7].

In part-of-speech tagging, the model processes input text word by word and learns the associations between words and their corresponding POS tags by adjusting the connection weights. Once trained, the neural network can assign POS tags to new text by utilizing its acquired knowledge through the learned connections. Neural network-based POS tagging surpasses traditional rule-based or statistical methods due to its capability to capture complex contextual information and achieve superior generalization on unseen data.

There are multiple variants and architectures available for neural network-based POS taggers [22] [56] [57] [58] [59] [60] [61]. Some common variants include Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Networks (CNN), BERT (Bidirectional Encoder Representations from Transformers), and combinations thereof.

1.5 Motivation

Part of speech tagging is one of the fundamental modules required for the development of a language in the computational field. A significant amount of information about a word and its surrounding words can be gleaned from its part of speech, which can then be incorporated into a language model and used for a

variety of speech and natural language processing applications. The performance of a POS tagger directly determines the quality and reliability of many of its subsequent phases of NLP tasks. It finds important roles in information retrieval, sentiment analysis, question and answering, machine translation, and many more NLP applications. So establishing a proper tagging system for the Mizo is an essential driving force for the development of the language.

With the advent of new technologies, research in the field of language processing has taken a giant leap forward, and a lot of researchers have given attention to the research field of computational linguistics. Numerous studies in part-of-speech tagging have already been conducted for Chinese, English and many European languages, and above 97% accuracy has already been achieved [62]. Lexical resources for many languages become stable and give open space for further development of the language in language processing.

In recent years, substantial efforts have been been put into the POS tagging for a variety of Indian languages, including Hindi, Urdu, Punjabi, Bengali, Gujarati, Kannada, and many other Indian languages. In addition, a significant amount of research is still being conducted on a variety of Indian languages. Despite all the advancements in computational linguistics fields for different languages, the Mizo language lags far behind other major Indian languages. When it comes to POS tagging for the Mizo language, it's important to note that despite its linguistically rich heritage, Mizo has received limited attention in the field of computational linguistics. There has been minimal research conducted in this specific area. This underrepresentation underscores the fact that Mizo has yet to fully tap into the vast potential of advancements in language processing, particularly in the realm of POS tagging.

Mizo is a grammatically rich and complex language with limited computational research tools. From a computational standpoint, the Mizo language presents numerous challenges. In fact, Mizo's diverse grammatical structures and distinctive characteristics make it even more appealing for researchers to work in

this language. It is expected to advance research in this language by developing more lexical tools.

To perform POS tagging, a large enough dataset is required to get an accurate result. Having a lot of resources that are free to use is one of the most crucial matters for the development of a language in the field of computing. Due to the Mizo language being under-resourced and yet in its infancy, the publicly available resources are minimal. The United Nations Education, Scientific and Cultural Organisation (UNESCO) has recognized the Mizo language as an endangered language [63]. In such circumstances, it is necessary to give a great deal of research attention and more effort to make the language develop with modern language technology.

1.6 Objectives

The main objective of this research work is to design and build a reasonably good accuracy part-of-speech tagger for the Mizo language. To accomplish this overarching goal, we have outlined the following key points:

- *Computational Analysis:* Conduct an in-depth computational analysis of the Mizo language to understand its complex behavior. This analysis will involve identifying patterns, linguistic features, and context-specific cues that can contribute to the development of an effective part-of-speech tagger.
- *Exploration of Approaches*: Explore and evaluate different methodologies, techniques, and frameworks employed in the development of part-of-speech tagging systems. This exploration will help identify the most suitable approaches for the unique characteristics of the Mizo language.
- *Resource Development:* Establish a reliable tagset and curate a substantial dataset of annotated corpora specifically for the Mizo language. This entails creating a comprehensive collection of labeled linguistic data that will serve as a valuable resource for training and evaluating the part-of-speech tagger.

These resources will facilitate further advancements in language processing applications for Mizo.

- *Grammar-based Insights:* Develop a precise and reliable regular expression framework that can provide meaningful insights into the grammatical properties and structures of Mizo words. These expressions will serve as valuable indicators, aiding the performance enhancement of the part-of-speech tagger by providing additional linguistic context and improving the accuracy of the tagging process.
- *Novel Tagger Design:* Devise and construct a novel part-of-speech tagger specifically tailored for the Mizo language. This innovative tagger design aims to introduce new approaches, techniques, or methodologies that go beyond existing solutions.

1.7 Methodology

The development of the Mizo part-of-speech tagger is comprised of multiple stages. The following points presented briefly in this section serve as a road map for the development process of the proposed tagger.

(i) Collecting Mizo Electronic Text: To ensure the tagger's accuracy in correctly labeling the parts of speech in Mizo, it is crucial to train it using meticulously annotated corpora. Unfortunately, publicly accessible corpora for Mizo language are limited. To overcome this challenge, we undertook an effort to collect a substantial corpus of Mizo electronic text. Our primary source of data was reliable online Mizo newspapers. By collating this data, we aimed to generate a sizable and representative corpus for training the tagger.

(ii) Pre-processing the Raw Data: Raw text data often contains noise, extraneous information, and inconsistencies that can hinder the accuracy and interpretability of the results. To mitigate these issues, it was essential to perform appropriate pre-processing on the collected raw Mizo text. This pre-processing

involved several steps, including cleaning noisy data, normalizing words affected by varying writing styles, and rectifying any spelling mistakes. Additionally, the sentences were tokenized into individual words, enabling further processing and analysis.

(iii) Designing a Well-Defined Tagset: A well-defined tagset plays a vital role in building an effective tagging system. It consists of a collection of tags or labels that indicate the morphological classes of each word in the sentences. To create an appropriate tagset, we conducted an in-depth study of the grammatical structure of the Mizo language. This analysis informed the development of a tagset that accurately signifies the grammatical information associated with each token in the corpus.

iv) Manual Tagging of the Corpus: To create a labeled dataset for training and evaluation, the collected corpus was manually tagged using the designed tagset. Each token in the corpus was annotated with the corresponding tag, indicating its morphological class. This tagged corpus was then divided into a training set and a testing set, allowing for the evaluation of the tagger's performance.

(v) Developing the Tagger: Based on the designed tagset and the manually tagged corpus, we developed Mizo part-of-speech tagger. These taggers utilize the tagset's labels to identify and assign appropriate tags to the tokens in the corpus. By employing advanced computational techniques and algorithms, the taggers aim to accurately label the parts of speech, contributing to the understanding and analysis of Mizo language texts.

(vi) Detailed Study and Comparison of Approaches: Utilizing the training dataset, we conducted an in-depth study and comparison of various approaches for developing the Mizo part-of-speech tagger. By exploring different methodologies, algorithms, and techniques, we aimed to identify the most effective approach that yields accurate and reliable results for tagging Mizo language texts.

By following this comprehensive methodology, we strive to develop robust and accurate part-of-speech taggers specifically tailored for the Mizo language. This endeavor addresses the challenges posed by limited publicly accessible corpora and incorporates linguistic insights to enhance the understanding and analysis of Mizo texts.

1.8 Thesis Organization

The subsequent sections of this thesis are structured into the following chapters:

Chapter 2 offers a concise review of the existing literature in the field of part-of-speech (POS) tagging. By examining prior research, this chapter establishes a foundation for understanding the various approaches, methodologies, and challenges associated with POS tagging across different languages and contexts.

Chapter 3 delves into the specifics of the Mizo language itself. This chapter provides an in-depth exploration of the linguistic characteristics, including syntactic structures and morphological intricacies, that distinguish Mizo. The purpose is to understand the unique challenges posed by the language, thus justifying the need for customized tagging methods.

Chapter 4 introduces the Hidden Markov Model (HMM) as a primary tagging approach. The theoretical underpinnings of HMMs in the context of POS tagging are explained, followed by an exploration of how these models are applied to linguistic analysis. This chapter documents the adaptation of HMMs to Mizo, encompassing data preprocessing, feature extraction, and model training. The subsequent evaluation quantifies the performance of the HMM-based POS tagger on Mizo text.

Chapter 5 introduces an alternative approach to POS tagging: the Conditional Random Fields (CRF) model. The adaptation of CRFs to the intricacies of the Mizo language, including feature engineering and model optimization, will

be detailed. This chapter culminates in the presentation of results and analysis of the CRF-based tagger's performance.

Chapter 6 introduces an innovative approach: the Hybrid Tagger. This chapter details the amalgamation of Hidden Markov Models, N-gram, and rule-based methods to enhance POS tagging accuracy for Mizo. The rationale behind this hybrid integration is explained, followed by a description of the practical implementation process. The chapter culminates in a comprehensive evaluation of the hybrid approach's efficacy and its potential benefits.

Chapter 7 serves as the conclusive chapter, summarizing the entire research endeavor. It provides a concise overview of the results obtained from each tagging approach and discusses their relevance to POS tagging in the Mizo language. The chapter also reflects on the attainment of research goals and explores the broader implications of the contributions made. Additionally, it concludes by outlining potential directions for future research, such as exploring advanced deep learning methods, expanding dataset sizes, and adapting methodologies for other languages with limited linguistic resources.

Finally, it is followed by a list of references at the end.

Chapter 2

LITERATURE SURVEY

The field of natural language processing (NLP) has witnessed significant advancements in recent years, with a particular focus on developing accurate and efficient part-of-speech (POS) taggers. POS tagging plays a crucial role in various NLP applications, such as syntactic parsing, machine translation, information extraction, and sentiment analysis. POS taggers enable a deeper understanding of the grammatical structure and semantic meaning of textual data, by assigning appropriate POS tags to each word in a sentence.

Numerous research have been conducted to enhance the performance and capabilities of POS taggers. These studies encompass a wide range of approaches, including rule-based methods, statistical models, and neural network-based techniques. Initially, researchers relied on manual rule engineering for POS tagging. Linguistic taggers incorporated knowledge in the form of rules or constraints crafted by linguists. However, in recent times, statistical and probabilistic models have gained popularity for their ability to provide adaptable and transportable taggers. These models leverage sophisticated machine learning algorithms to acquire robust information. However, most statistical models depend on manually POS-labeled corpora to learn the underlying language model, which can be challenging to obtain for new languages.

This chapter aims to provide a brief overview of prior works in POS tagging, with a particular focus on the different techniques employed in this field. While not exhaustive, this review will shed light on the key approaches utilized in POS tagging. Additionally, this chapter will delve into a detailed review of POS taggers developed specifically for Indian languages.

Through this review, we aim to provide a comprehensive understanding of

the existing research on POS taggers, their methodologies, and the current state-of-the-art. This knowledge will lay the groundwork for the subsequent section, where we will present a detailed analysis of POS tagger related works specific to Indian languages.

2.1 Corpus Repository

Part-of-speech tagging, an essential component of natural language processing, is closely intertwined with corpus linguistics. Numerous publicly accessible POS corpora have been created, while certain ones may entail payment for access. Below are notable instances:

Brown Corpus: A significant milestone in English corpus development for computer analysis was the creation of the Brown Corpus at Brown University during the mid-1960s [64]. This comprehensive corpus, created by W. Nelson Francis and Henry Kučera, consists of approximately one million words. It encompasses 500 samples that were randomly selected from a diverse range of publications, making it a valuable resource for in-depth studies of the English language.

Penn Treebank (PTB): The Penn Treebank is one of the most widely used treebanks for POS tagging and other linguistic tasks.The project started in the mid-1980s and resulted in the creation of a large annotated corpus for English, including POS tags [65]. It contains annotated data from various genres, including newswire, text from the Wall Street Journal (WSJ), and other sources.

Universal Dependencies (UD): This project offers POS-tagged corpora for numerous languages, promoting cross-linguistic research in NLP. It aims to establish consistent annotation standards across different languages. The initial release of the dataset occurred in 2015 and included 10 treebanks covering 10 languages [66]. In the 2020 release, version 2.7, the dataset expanded to encompass 183 treebanks across 104 languages. The annotation in these treebanks includes universal part-of-speech tags (UPOS), language-specific part-of-speech tags (XPOS), uni-

versal morphological features (Feats), lemmas, dependency heads, and universal dependency labels.

The American National Corpus (ANC): It is a collection of 22 million words of written and spoken American English data since 1990 [67]. It includes diverse genres like email, tweets, and web data. The ANC is annotated for part of speech, lemma, shallow parse, and named entities. A subset called the Open American National Corpus (OANC) is freely available with no restrictions on its use from the ANC Website.

The Google Universal POS Tags: Petrov et al. [68] developed a universal part-of-speech (POS) tagset with twelve categories and provided a mapping from 25 language-specific tagsets to this universal set. Using this tagset and mapping, they created a standardized dataset with POS annotations for 22 different languages. The tagset and mappings can be downloaded from `http://code.google.com/p/universal-pos-tags/`.

OntoNotes: OntoNotes [69] is a project that annotated a diverse corpus of text genres in English, Chinese, and Arabic with structural and semantic information. OntoNotes Release 5.0 version was the final release, created through collaboration between BBN Technologies, the University of Colorado, the University of Pennsylvania, and the University of Southern California's Information Sciences Institute. It provides a large multilingual corpus with rich annotations, reaching 90% interannotator agreement.

The Linguistic Data Consortium for Indian Languages (LDC-IL): The Linguistic Data Consortium for Indian Languages (LDC-IL) is a vital repository for linguistic resources in all Indian languages, including text, speech, and lexical corpora [70]. It was established in 2007, under the Department of Higher Education, Ministry of Human Resource and Development, Government of India. Housed within the Central Institute of Indian Languages in Mysore, it aims to become a self-sufficient institution by developing and distributing linguistic resources to developers, researchers, and organizations. LDC-IL has been dis-

tributing linguistic resources for AI and NLP in Indian languages since April 4, 2019 through its Data Distribution Portal.

The presence of numerous corpora provides valuable resources for the study and enhancement of part-of-speech tagging in diverse languages and domains. Researchers and practitioners can take advantage of these corpora to create and assess robust and precise POS taggers. The wide range of linguistic structures and genres covered by these corpora presents opportunities to explore innovative techniques and approaches, effectively addressing the challenges of POS tagging.

2.2 Part-of-speech Taggers

In the realm of POS tagging, the 1960s marked the beginning of significant advancements [71][72][37][73][74]. Since then, researchers have continuously sought to enhance the accuracy and efficiency of POS tagging techniques across various languages. In the beginning, rule-based or stochastic methods were predominantly utilized in numerous tagging systems.

2.2.1 Rule-based Tagging Systems

TAGGIT [74], an influential rule-based POS tagger, emerged as one of the pioneering works in the field. It was primarily used for tagging the Brown Corpus and achieved a commendable accuracy rate of 77%. However, the remaining portions of the corpus required manual tagging efforts spanning several years. Another notable figure in rule-based POS tagging is Eric Brill [32], who developed a widely adopted tagger. The paper reported achieving accuracy rates of around 96% on certain datasets, showcasing the effectiveness of the rule-based approach.

In the mid-1990s, the ENGTWOL tagger, proposed by Karlsson et., [75], made significant strides as another noteworthy rule-based approach for English POS tagging. The ENGTWOL aimed to enhance the accuracy and efficiency of

the tagging process, contributing to the advancement of rule-based POS tagging methodologies.

Aone et al. [33] described a successful implementation and extension of Brill's unsupervised learning algorithm for a Spanish Part-of-Speech (POS) tagger. The implementation employed a Spanish lexicon and morphological analyzer to alleviate ambiguity in POS tagging when processing Spanish texts. The system achieved approximately 92% accuracy, with potential for further improvement by addressing existing challenges and including unknown words.

Sharipov et al. [76] introduced a part-of-speech (POS) annotated dataset and rule-based POS-tagger tool for the low-resource Uzbek language. The dataset covered various fields and employs an affix/suffix stripping approach for stemming. The tagger tool demonstrated high accuracy, achieving approximately 90% overall accuracy when tested on the annotated dataset with more than 20 fields. This dataset and tagger tool provided valuable resources for natural language processing tasks in Uzbek and related Turkic languages, offering a solid foundation for future NLP advancements in the language.

A rule-based part-of-speech (POS) tagger for the Indonesian language was presented by Rahel et al. [77]. The system incorporated tokenization, multi-word expression handling, and named entity recognition. It assigned tags to each token, following a rule-based approach that disambiguates the tags. The system achieved an accuracy of 79% on a manually tagged corpus of approximately 250,000 tokens. The work provided valuable insights into POS tagging for Indonesian and served as a foundation for further improvements in accuracy and performance.

In their study, Purnamasari et al. [34] proposed a novel approach to part-of-speech (POS) tagging in Indonesian that did not rely on annotated corpora. Instead, they utilized the Indonesian dictionary (KBBI) and morphological rules to handle word form changes. Their rule-based POS tagger achieved an impressive accuracy rate of 87.4% on the PAN Localization corpus for Indonesian.

In the paper by Bao Pham [35], a rule-based part-of-speech (POS) tagger was developed for the English language using Lex and Yacc. The tagger employed a small set of rules and a dictionary to generate token sequences. The author highlighted the efficiency and speed of rule-based tagging due to its linguistic rules. However, both rule-based and stochastic approaches faced challenges with ambiguity and unknown words. The rule-based approach excelled in tagging unknown words in languages with rich morphology, while stochastic approaches required more time but offered higher accuracy for ambiguous or unknown words. The paper provided valuable insights into the trade-offs between rule-based and stochastic approaches for POS tagging.

In their paper, Vaishali et al. [36] presented a part-of-speech (POS) tagger specifically designed for Marathi, a widely spoken Indian language. The system adopts a rule-based approach that leverages Marathi transformational grammar to accurately assign part-of-speech tags to Marathi words. Manual annotation of a corpus is conducted to tag words with their respective POS. The proposed system achieved an impressive overall accuracy of 97.56%, showcasing its effectiveness for preprocessing tasks and natural language processing (NLP) applications in Marathi.

Alex et al. [78] presented a POS tagger for the Kadazan language, utilizing Brill's approach in their paper. This tagger automated tagging, reduced ambiguity, and achieved an impressive accuracy of approximately 93%. This study addressed the lack of a POS tagger for the Kadazan language, demonstrating the effectiveness of Brill's approach. It emphasized the tagger's potential for language learning and NLP applications in the context of the Kadazan language.

Roche et. al [79] proposed a novel approach to part-of-speech tagging using finite-state transducers. They addressed the slow execution of rule-based taggers and achieved optimal time efficiency by transforming them into deterministic transducers. The paper lacked empirical evaluation, but the approach showed potential for improving part-of-speech tagging.

2.2.2 Stochastic-based Tagging Systems

Stochastic POS tagging approaches utilize statistical models to assign part-of-speech tags. In response to the limitations of rule-based taggers, researchers introduced statistical and probabilistic models as a solution. The stochastic approach offered a more adaptable and data-driven method for assigning part-of-speech tags to words in a given text. Early pioneers in this field, such as [71], [37], and [73], were instrumental in developing and implementing stochastic taggers. Their contributions laid the groundwork for subsequent advancements and enhancements in the performance of tagging systems.

The English Penn Treebank project [65] stands out as a significant milestone in the field of annotated corpora. One notable breakthrough within the stochastic modeling approach was Thorstens Brants' proposal of Trigrams'n'Tag (TNT) [80]. Brants introduced an efficient statistical POS tagger that leveraged second-order Markov models. The TNT tagger demonstrated superior accuracy compared to its counterparts while also boasting remarkable speed in both training and tagging processes. Brants' work highlighted the potential for combining statistical modeling techniques with POS tagging, paving the way for further advancements in the field. The TNT tagger achieved an average accuracy of 96% to 97%, depending on the language and tagset.

Dandapat et al. [81] introduced a Bengali Part-of-speech tagging (POS) approach, utilizing a Hidden Markov Model (HMM)-based stochastic tagger that incorporates morphological and contextual information. To enhance results given the limited labeled training set (41,000 words), they incorporated a morphological analyzer and implemented semi-supervised learning using a larger unlabeled training set (100,000 words). The tagger achieved an accuracy of 89% on the provided test data.

Kupiec et al. [38] proposed a part-of-speech tagging system based on a hidden Markov model. Techniques like word equivalence classes and selective extension

of context via predefined networks ensure robustness with high performance. Unknown word categories are predicted using local context and suffix information. Evaluation using the Brown corpus shows an impressive 96% accuracy. The system's versatility is exemplified in its successful application to French tagging

Lee et al. [82] presented uniformly lexicalized HMMs for POS tagging in English and Korean, effectively addressing data sparseness with simplified backoff smoothing. Their lexicalized models outperformed non-lexicalized ones, achieving error reduction ratios of 24.20% (English) and 39.95% (Korean). They are now exploring the transformation of their uniform models into non-uniform ones, aiming for better space complexity and reliable parameter estimation without sacrificing accuracy.

Cing et al. [83] presented joint word segmentation and POS tagging in Myanmar using HMM and morphological rules. The proposed approach outperformed separate word segmentation and POS tagging methods. Significant improvement was observed with the joint approach using HMM and morphological rules.

Huda et al. [84] explored Arabic Part of Speech (POS) tagging using the Hidden Markov Model (HMM) method on Qur'an text. Their dataset comprised 150 simple sentences, 50 compound sentences, and 50 verses representing complex sentences from the Qur'an corpus. They utilized K-Fold Cross Validation for experimentation and obtained promising average accuracies of 89.44% for simple sentences, 74.18% for compound sentences, and 69.04% for complex sentences.

Adafre [85] used Conditional Random Fields (CRFs) for Amharic word segmentation and POS tagging with a small annotated corpus of 1000 words. Despite the limited data size and a large number of unknown words (80%) in the test corpus, the results were encouraging. They achieved an accuracy of 84% for Amharic word segmentation and 74% for POS tagging, demonstrating the applicability of CRFs for handling the complexities of a morphologically rich language like Amharic.

Pisceldo et al. [86] introduced a probabilistic approach for building Part-Of-Speech (POS) taggers for Bahasa Indonesia. They utilized two methods, Maximum Entropy and Conditional Random Fields (CRF), for tagging terms in their study. By comparing the two approaches using data from different corpora, they found that the Maximum Entropy approach yielded the best results, achieving an accuracy of 97.57%. The CRF approach also demonstrated a competitive accuracy of 90.46% in their experiments.

Silfverberg et al. [87] focused on enhancing the accuracy of CRF-based POS tagging by leveraging sub-label dependency structure. They incorporated the dependencies into the CRF model through a straightforward feature expansion scheme. Through experiments conducted on five different languages, the researchers demonstrated that this approach could lead to a substantial improvement in tagging accuracy, particularly when using fine-grained label sets.

Fanoon et al. [88]developed a specialized part-of-speech tagging system for Twitter text data using the CRF toolkit. The approach aimed to address the unique challenges of Twitter conversations. The system was trained on nearly 1000 Twitter conversations and showed significant efficiency in tagging accuracy.

Huang et al. [89] introduced an ME-based POS tagging model, which outperformed HMMs. ME models can leverage non-independent features, benefiting POS tagging without considering dependency distributions. The flexibility of ME models efficiently handles sparse training data. Experimental results showed a substantial error rate reduction compared to the HMM-based baseline, achieving 98.01% accuracy in close test and 95.56% in open test.

Cahyani et al. [90] developed a POS tagging system to address word ambiguity in Indonesian text. They used the maximum entropy Markov model (MEMM) on a manually labeled "Indonesian manually tagged corpus" and obtained average accuracies of 83.04% for the MEMM Bigram algorithm and 86.66% for the MEMM Trigram algorithm, showing the superiority of MEMM Trigram over previous methods. introduced a Chinese POS tagger based on the maximum entropy

model. Trained on a large corpus, it accurately predicted POS tags, achieving an impressive 96.8% accuracy in open testing for Chinese POS tagging.

Yi et al. [91] utilized the maximum entropy model for English part-of-speech tagging, enhancing it with pre-tagging to include additional context features. They also improved the tagging algorithm, optimizing the POS series without extra computation, leading to improved accuracy. The combined algorithm achieved 94% accuracy and recall rate, fully leveraging the advantages of the maximum entropy method

2.2.3 Other Approaches

In addition to rule-based and stochastic tagging approaches, there are other notable classifications in POS tagging, including Hybrid POS tagging and Deep Learning methods, Genetic Expression Programming which have gained popularity in recent years.

Kassanhun et al. [92] proposed a deep learning-based POS tagging method for Ge'ez language. Using sentences from the Holy Bible, they train GRU, Bi-GRU, LSTM, and Bi-LSTM models with varying hyperparameters. The best-performing model, Bi-GRU, achieved 86.70% accuracy with 10 epochs, 2 hidden layers, 128 neurons per layer, and a learning rate of 0.01. This research highlighted the effectiveness of deep learning for Ge'ez POS tagging despite the absence of a standardized corpus.

Dalai et al. [93] introduced solutions to overcome the scarcity of resources and tools for Odia POS tagging. They presented CRF and deep learning models, achieving accuracies of 92.08% and 94.58% respectively, with the Bi-LSTM model incorporating CNN and CRF layers. These approaches surpassed previous studies in Odia POS tagging.

Xu et al. [94] proposed a data-driven model for automated Chinese word segmentation and POS tagging. Their approach incorporated word2Vec for fea-

ture extraction, a modified AlexNet for further feature extraction, and a neural network language model in the hidden layer. The output layer employed a conditional random field and an auxiliary loss function to improve training. Experimental results showed superior performance in both micro-F1 and macro-F1 metrics compared to other models, while maintaining high operational efficiency.

Liu et al. [95] proposed a uniform-design genetic expression programming (UGEP) model for POS tagging. The UGEP model outperforms genetic algorithm models, neural networks, and hidden Markov models (HMMs) on the Brown Corpus. It achieved a high accuracy rate of 98.8% in closed lexicon tests and 97.4% in open lexicon tests, with 88.6% accuracy on unknown words.

Lv et al. [96] introduced a Genetic Expression Programming (GEP) model designed for POS tagging. The experimental findings on the Brown Corpus revealed that the proposed GEP model outperforms both Genetic Algorithm models and HMM models, achieving a significantly higher accuracy rate of 97.40% in POS tagging tasks.

Xue et al. [97] proposed a specialized POS tagger for building codes to enable automated regulatory rule conversion. Their deep learning neural network model, fine-tuned on Part-of-Speech Tagged Building Codes (PTBC) dataset, achieved 91.89% precision without error-driven rules and 95.11% precision with rules, outperforming state-of-the-art POS taggers at 89.82%. Imani et al. [98] propose a novel label transfer method for low-resource languages, leveraging graph-based label propagation and a Graph Neural Network with transformer layers. Their approach achieved state-of-the-art results in unsupervised POS tagging when combined with enhanced contextualized embeddings.

Reyes et al. [40] explored the utilization of Support Vector Machines (SVM) in POS tagging for Filipino language. The approach demonstrated promising potential by achieving an impressive accuracy rating of 81%, surpassing the performance of existing POS taggers for the Filipino language by 2%. Murata et al. [41] introduced new tagging methods using decision-list, maximum entropy, and

support vector machine approaches. Their experiments revealed that the support vector machine method achieves the highest precision (96.1%) and significantly improves tagging accuracy for the Thai language.

Dibitso et al. [51] investigated methods to improve tagger accuracy for Setswana, a language with limited resources. They utilize Maximum Entropy (MaxEnt) and Support Vector Machine (SVM) taggers, achieving accuracies of 94.4% and 95.59% respectively. By combining these taggers using a voting algorithm, they achieved an improved accuracy of 97.06%.

Besharati et al. [99] introduced a hybrid model that combines HMM and a single-layer bidirectional LSTM for Persian POS tagging. This innovative approach improved the accuracy of both HMM and neural models, achieving an accuracy of 97.29%. Saidi et al. [61] presented a novel grammatical tagging system for Arabic language using BERT, a large pre-trained language model. Despite the limited availability of BERT-based models in Arabic, their system achieved high accuracy of 91.69% in labeling input sentences with the most likely sequence of tags. They constructed a substantial corpus by combining existing corpora like Arabic WordNet and the Quranic Arabic Corpus.

2.3 Indian Languages POS Tagging

A considerable amount of research work on POS tagging for different Indian languages has already been done, and a selection of these studies is presented below.

Among the initial works in the field of POS tagging for the Hindi language [100], [101], and [102] stand out as notable contributions. Sarkar et al. [103] developed a Trigram HMM-Based POS Tagger for Indian Languages, demonstrating its superior performance compared to a bigram tagger across Hindi, Bengali, Marathi, and Telegu. Their experiment emphasized the effectiveness of the trigram tagger's analysis of prefix, suffix, and word-type in handling unknown words

over a suffix-only approach. Joshi et al. [104] proposed Hidden Markov Model based POS tagger for Hindi utilizing the IL POS tagset. They developed a test corpus consisting of 500 sentences (11,720 words) and reported a system accuracy of 92.13%.

Sharma et al. [39] proposed a POS tagger for the Panjabi language, employing a bigram Hidden Markov model implemented with the Viterbi algorithm. Their model was evaluated on a corpus containing 26,479 words, achieving an accuracy of 90.11%. Gupta et al. [105] developed an automatic POS tagger for Urdu, utilizing both HMM and Conditional Random Field (CRF) approaches. They used 7,000 sentences as training data for machine learning and obtained accuracies of 83.37% and 81.03% on the test dataset using CRF and HMM-based approaches, respectively.

Patra et al. [106] described the development of a POS tagger using rule based and supervised methods in Kokborok, a resource constrained and less computerized Indian language. They utilized 26 tagsets to tag a corpus of 42,537 words. In the statistical methods, they employed two machine learning classifiers, Support Vector Machines (SVM) and Conditional Random Field (CRF).They reported accuracies of 69% for rule-based, 81.67% for CRF-based and 84.46% for SVM-based methods.

Bharti et al. [107] proposed a heuristic-based approach for POS information identification. Their method utilized a context-based bigram model to determine the most likely POS information for each word based on adjacent word relationships. The approach outperformed existing techniques, achieving 94.3% accuracy in POS tagging for Hindi.

Deskmukh et al. [108] addressed the need for accurate part-of-speech (POS) tagging in Marathi, a language with limited NLP tools and corpora. They proposed two models: a deep learning model and a bidirectional long short-term memory (Bi-LSTM) model for POS tagging in Marathi text. The deep learning model achieved 85% accuracy, while the Bi-LSTM model reached an impressive

97% accuracy.

Shree et al. [109] developed a deep neural network-based POS tagger model for Kannada. The model combined word embedding, RNN, and LSTM techniques. The POS tagger achieved 81% accuracy on unseen data from a dataset of 10,000 annotated Kannada sentences. This research made a significant contribution to the field of computational linguistics in the context of Kannada.

Tailor et al. [52] presented a hybrid approach to enhance the accuracy of POS tagging for the Gujarati language. The proposed method combined LSTM-based POS tagging with Computational Linguistic Rules. By incorporating language-specific rules, the accuracy of the statistical taggers was significantly improved.

Akhil et al. [110] proposed a deep learning-based method for POS tagging in Malayalam. Through the training of four deep learning architectures on a publicly available tagged Malayalam dataset, they showcased the superiority of their approach over existing state-of-the-art methods in language computing tasks, especially in POS tagging for Indic languages. This research significantly contributes to the progress of POS tagging techniques, with a particular emphasis on low-resource language processing.

In their study, Advaid et al [111] focused on the development of a POS tagger for Kannada and Hindi languages using ML and DL algorithms. Through experiments conducted on a combined corpus of around 300,000 unique words, they demonstrate the effectiveness of their approach. The proposed POS tagger utilized a set of 17 tags from the BIS tag set, successfully addressing the challenges associated with POS tagging in Kannada and Hindi.

Ovi et al. [56] introduced a novel multi-phase recurrent neural network (RNN), for Bangla Parts-of-Speech tagging, named BeNeP. BaNeP combined a bidirectional LSTM-based sub-network to extract structural features with a weighted context extraction procedure that captures intricate contextual relations between words in a sentence. Experimental results on the LDC2010T16 dataset show-

case the effectiveness of BaNeP, surpassing existing Bangla POS taggers with a significant improvement in accuracy.

Mundotiya et al. [112] proposed an attention-based model with self-attention and monotonic chunk-wise attention to leverage syntactic relations even with limited annotated data from the Hindi Disease domain. The attention-based model achieved an accuracy of 93.86%, showing improvement over the baseline model (93.64%). However, the baseline model outperforms the attention model in terms of F1-score, with 93.65% for the baseline and 94.05% for the attention model.

Rajan et al. [113] presented a pioneering experiment in the application of deep learning techniques for part-of-speech (PoS) tagging in the Konkani language. By utilizing a dataset of over 100,000 PoS-tagged Konkani sentences, the authors claimed that their approach outperformed previous studies with f-scores of 90.73%. Sathsarani's research [114] introduced a novel deep learning-based approach for POS tagging in Sinhala, overcoming the challenges of low-resource language processing. By combining highly accurate individual classifiers for primary POS tags into a composite model, the proposed solution outperforms traditional methods in terms of accuracy.

Tehseen et al. [57] presented the development and evaluation of a POS tagged corpus and a BiLSTM-based POS tagger for Shahmukhi (Western Punjabi). The corpus consists of 0.13 million words from 14 different domains, and the tagger demonstrated remarkable performance, achieving an f-score of 96.11% and an accuracy of 96.12%.

Anbukkarasi et al.[115] employed a range of deep learning models, including RNN, LSTM, GRU, and Bi-LSTM, for word-level POS tagging in Tamil. The training process utilized a tag set consisting of 32 tags and a corpus of 225,000 tagged Tamil words. The experimental findings revealed that enhancing the hidden state led to improved model performance, with the Bi-LSTM model achieving the highest accuracy of 94%.

Dutta et al.[116] presented an Intelligent POS tagger for Hindi language that incorporated Viterbi and K-Nearest Neighbour algorithms. The proposed tagger outperformed Viterbi, especially when dealing with unknown words, resulting in improved accuracy. Warjri et al. [48] employed deep learning techniques to create a POS tagger for the Khasi language. The taggers were evaluated on the designed Khasi corpus, yielding impressive accuracies of 96.81% for BiLSTM, 96.98% for BiLSTM with CRF, and 95.86% for character-based LSTM.

Pakray et al. [117] presented a pioneering effort in their paper as the first known endeavor in the domain of part-of-speech tagging for the Mizo language. To the best of our knowledge, this work stands as the sole and initial contribution to the field of POS tagging for the Mizo language. Their work encompasses two key facets: the establishment of a Mizo-to-English dictionary and the development of a part-of-speech tagger. Notably, the Mizo-to-English dictionary consists of a substantial compilation of 26,407 entries, meticulously curated through a combination of manual and automated processes. This groundbreaking paper commences with an exploration of Mizo parts of speech, forming the basis for constructing a specialized part-of-speech tag list. Remarkably, the authors introduce a novel 24-item tag set, purposefully designed to augment the precision and applicability of part-of-speech tagging for the Mizo language. With their dictionary and part-of-speech tag set as cornerstones, the authors lay the groundwork for the creation of an automatic part-of-speech tagger tailored to the unique attributes of the Mizo language. This seminal work not only fills a significant gap but also paves the way for further advancements in linguistic analysis and technological applications for the Mizo language.

This literature review highlighted the evolution of taggers from early rule-based methods to the more recent advances in stochastic models and deep learning techniques. Researchers have explored diverse linguistic resources, annotated corpora, and innovative algorithms to improve tagging accuracy and performance. The thorough examination of various POS tagging systems across languages, including English, Chinese, European languages, Indian languages, and numer-

ous others, highlights their impressive accuracy and performance. However, the scarcity of literature on the POS tagging of the Mizo language underscores the need for further research and development tailored to languages with limited resources.

The reviewed literature also emphasizes the importance of accounting for language-specific variations and challenges. While existing systems excel in many linguistic contexts, addressing the unique characteristics of languages like Mizo is crucial for achieving accurate POS tagging. This study identifies a compelling opportunity for future research to bridge the gap in POS tagging for underrepresented languages, enriching the landscape of natural language processing and enabling more effective applications in diverse linguistic settings. In the following chapters of this thesis, we will investigate the morphological structure of the Mizo language and discuss the practical measures taken to further enhance the development of a POS tagging system tailored to the Mizo language.

Chapter 3

MIZO LANGUAGE

Studying and exploring the Mizo language, along with its complex elements, carries great significance in the field of part-of-speech (POS) tagging research related to the Mizo language. POS tagging involves assigning grammatical labels to individual words in a sentence. By gaining a thorough comprehension of the parts of speech and the unique linguistic features specific to the Mizo language, researchers can develop more precise and efficient POS tagging models.

In this chapter, our aim is to provide an extensive introduction to the Mizo language. We will delve into its historical context, investigate the various parts of speech that exist within the language, delve into its distinct linguistic characteristics, and carefully examine the computational challenges that arise when working with it. By doing so, we can contribute to the development of more sophisticated computational tools and methods for effectively processing and analyzing the Mizo language.

3.1 Overview

There are numerous numbers of languages in the world, spoken by human beings. These languages are very complex, diverse, and unique on their own. According to the 25^{th} Edition of the Ethnologue[118], there are around 7151 living human languages in the world today, out of which English has the most speakers, with around 1452 million speakers from different 146 countries. Whereas considering native speakers only, Mandarin Chinese has the highest number in the world, with 920 million native speakers followed by Spanish, with around 475 million native speakers. Of the world's 142 language families, Trans-New Guinea, Niger-Congo,

Indo-European, Austronesian, Sino-Tibetan and Afro-Asian are the main families of languages making up five-sixths of the world's population.

Mizo language is one of the 456 languages in India [118], spoken mainly by the people of Mizoram, which is one of the 28 states in India. Other than the mainland of Mizoram, this language is also spoken in Myanmar, Bangladesh, and other parts of India, such as Tripura, Meghalaya, Assam, Manipur, and Nagaland. According to the 2011 Census, there are around 8.3 lakh people speaking the Mizo language in India and 8.45 lakh users all over the world. The language is categorized as an endangered language by the United Nations Educational, Scientific and Cultural Organization (UNESCO) [63].

The name of the state is derived from *"Mizo"*, the self-described name of the native inhabitants, and *"Ram"*, which in the Mizo language means "land". Thus "Mizoram" means "land of the Mizos". The nomenclature *'Mizo'* signifies both the name of a tribe and the language. The Mizo tribe is an ethnic group, settled mainly in the mountainous region of Mizoram and is believed to have migrated from China [119]. Other than mainland Mizoram, many of the Mizos reside in Myanmar, Tripura, Assam, Manipur, and Nagaland. The term Mizo is actually an umbrella term to denote the various clans, such as Hmar, Pawi, Paite, Lakher, Ralte, etc. Mizo as a language, belongs to the Tibeto-Burman group of languages[120] now spoken by around a hundred thousand individuals mostly from the territory of Mizoram and Chin State in Myanmar. The Mizo language, known as *Mizo ṭawng* or *Zo ṭawng* is also called as *Duhlian ṭawng*. The Mizo language, as in its present form, is generally based on Lusei(Lushai or Lushei) dialect but is evolved gradually with significant influence from its other Mizo sub-tribes such as Pawi, Hmar, Paite, Lai etc.

3.2 Historical Background

Thomas Herbert Lewin, fondly referred to by the Mizo as Thangliana, the then-Deputy Commissioner of the Chittagong Hill Tracts, deserves credit for the formalization of the Mizo language in Roman letters based on the Hunterian method of spelling. One of his books, published in the year 1874, named "Colloquial Exercises in the Lushai Dialect of the 'Dzo' or Kuki Language with vocabularies and Popular Tales", was considered the first significant attempt to write Mizo[121]. Lewin made an attempt to write a good number of Mizo words using the English alphabet in his book, which also included three folk tales[122]. Since then, he has since released a few helpful and informative books in the Mizo language's Lushai dialect. His contributions provided opportunities for others to study the Mizo language, which will contribute to its continued development in the future.

Later, in the year 1884, Brojo Nath Shaha, a Civil Medical Officer from Chittagong, published a Grammar of the Lushai Language, a scholarly work detailing the grammar of the Lushai language [123]. In the year 1885, a British Officer named C.A. Soppit published Rangkhol-Kuki-Lushai grammar, along with a variety of other helpful and relevant materials. All of these initiatives paved the way for the evolution of the Mizo language.

Development of the language took a significant leap forward when two Arthington missionaries Rev. F.W. Savidge and Rev. J.H. Lorrain, arrived in Aizawl on January 11, 1984 [119]. When they arrived, those pioneering written works on the Mizo language were extremely useful for them in their endeavors to learn Mizo words and phrases. They put lots of effort into reducing the Mizo language to writing for the native speakers, eventually settling on the Roman alphabet after adopting the then-current phonetic spelling method known as the Hunterian system of orthography. Savidge and Lorrain initially came up with the Mizo alphabet, which consists of 28 letters, as shown below [121]:

a, â, aw, âw, b, ch, d, e, ê, f, g, h, i, î, k, l, m, n, o, p, r, s, t, ṭ, u, û, v, z.

However, it was eventually discovered that some of the letters were confounding native speakers, and as a result, the presently accepted Mizo alphabet, which contains 25 letters as shown below, was devised after modifications [121].

a, aw, b, ch, d, e, f, g, ng, h, i, j, k, l, m, n, o, p, r, s, t, ṭ, u, v, z.

The majority of the letters are identical to the English letters, except for one letter T/t (big form/small form), printed with a subscript (.) (i.e., Ṭ/ṭ). All Mizo alphabets have a big letter form (capital) and a small letter form, similar to the conventional writing manner of the English alphabet. There are three letters, such as AW (aw), CH (ch), NG (ng), each of which consists of a cluster of two letters.

Not only did the two missionaries standardize and develop the Mizo alphabet, but they also translated a substantial section of the Bible into the Mizo language. With a commitment of over three years, they laid the groundwork for future systematic documentation of information about the regional language. For use in elementary schools, they created a number of books that beginners could use to learn how to read and write Lushai in Roman script.

3.3 Parts of Speech in Mizo language

There are several books on Mizo grammar that have been produced [120], [124], [125], [126], [127]. However, each book has its own unique breakdown of the parts of speech. In this section, the sub-classification of the Mizo parts of speech that are normally included in the majority of Mizo grammar books is examined.

Noun: In general, nouns are the words used to identify places, persons, animals, ideas, and things. The sub-classification of nouns in Mizo is discussed below:

- *Common Noun*: It refers to any generic name of things, i.e., common names given to a person, place, or thing. E.g., *Pangpar*(flower), *sava*(birds).

- *Proper Noun*: A proper noun is a noun that designates a specific entity, such as a person, place, or thing. In the Mizo language, proper nouns' first letter or all letters are usually capitalized. E.g., *Aizawl, Mawii.*

- *Abstract Noun*: An intangible concept such as an emotion, a feeling, a quality, or an idea. E.g., *beiseina*(hope), *hlauhna*(Fear).

- *Collective noun*: A collective noun is a name or noun that refers to a group of people, things, or animals. E.g., *mipui*(people), *zaipawl*(choir).

- *Material Noun*: Material nouns are names used to refer to materials or substances. E.g., *tui*(water), *tuboh*(hammer).

  The Mizo grammar experts Zarzova and Zikpuia [126] [127] introduced the following crucial additional sub-categories for Mizo:

- *Concrete Noun*: A concrete noun is one that can be tasted, touched, seen, heard, or smelled.

- *Countable Noun*: Countable nouns denote things that can be counted.

- *Uncountable Noun*: Uncountable nouns denote things that cannot be counted.

Pronoun: In general, a term substituted for a noun or a noun phrase in a sentence is known as a pronoun. It shortens and clarifies sentences by replacing nouns. Pronouns in Mizo are further classified as shown below:

- *Personal Pronoun*: Simple substitute for the proper name of a person or people. E.g., *Ka*(I), *Kan*(we).

- *Possessive Pronoun*: A pronoun that replaces a noun and shows ownership. e.g. *Ka*(my), *Kan*(Our).

- *Demonstrative Pronoun*: A pronoun that points towards the noun it replaces, indicating its position. E.g., *Saw saw, khi khi.*

- *Relative Pronoun*: A pronoun used to connect a clause or phrase to a noun or pronoun. e.g. *hi, chu, kha.*

- *Interrogative pronoun*: A pronoun used to ask questions is called an interrogative pronoun. E.g., *tu*(Who), *eng*(What)

  The following sub-classification is included in the grammar books [126] [127] published recently:

- *Emphatic pronoun*: An emphatic pronoun is a pronoun that comes right after the noun it refers to and is used for emphasis. E.g., *Keimah ngei*(myself), *nangmah ngei*(yourself).

Verb: Verbs are words that indicate an activity being carried out by the noun or subject of a sentence. It describes an activity or state of being. The Mizo language has the following sub-categories:

- *Transitive verb*: A transitive verb is a type of verb that requires an object to fully express the action that the subject is performing. E.g., *chhiar*(read), *pet*(kick).

- *Intransitive verb*: It is a verb that doesn't have a direct object yet signifies a complete action. E.g. *mu*(sleep), *kal*(go).

- *Double verb*: It is a verb that is written twice to show that the action is repeated. E.g., *au* means 'to shout'. *Ziak ziak* means 'to shout again and again.'

  The two most recent grammar books[126][127] expanded the following list of subcategories:

- *Nounal Verb*: Nounal Verbs are nouns that can act as a verb. E.g., *Ka Aizawl dawn e.*(I will go to Aizawl)

- *Adjectival Verb*: These are adjectives that play the role of the main verb in our sentence, E.g., *Kan uite chu a fing hle.*

- *Reflexive Verb:* A transitive verb is considered reflexive if both its subject and object always refer to the same thing or person. E.g., *inbual, inpeih.* Additionally, Zikpuia[127], has presented a more detailed taxonomy including reactionary verbs, reciprocal verbs, co-active verbs, causative verbs, exclusive verbs, common verbs, and phrasal verbs.

Adjective: Adjectives are words that are employed to further define or elucidate a noun or pronoun in a sentence. Below are the various sub-categories of adjectives in the Mizo language:

- *Adjectives of quality*: Adjectives of quality are those adjectives that describe the quality of a thing being described. E.g., *Mawi*(beautiful), *lian*(big)
- *Adjective of quantity*: Adjectives of quantity are those adjectives that describe the amount of a thing being described. E.g., *Tlem*(a few), *pakhat*(one)
- *Demonstrative Adjective*: Demonstrative Adjectives are those adjectives that describe the location of a thing being described. E.g., *Saw nula saw.*
- *Nounal Adjective*: It is a noun, that also plays the role of an adjective in the sentence. E.g., *Vawk sa*(pork).

  Zikpuia[127] further listed the following sub-categories of adjectives in the Mizo language: double adjective, Comparative adjective, superlative adjective, Interrogative adjective, emphatic adjective, the adjective of cause, intensifier adjective.

Adverb: Adverbs are words that are used to describe verbs, adjectives, and other adverbs in a sentence. The Mizo language has a number of different sub-categories of adverbs, which are shown below:

- *Adverb of Manner:* Adverbs of manner are adverbs that describe the manner in which something is done or occurs, E.g., *chak*(fast)

- *Adverb of Time:* Adverbs of Time specify the time of the action. E.g., *naktukah*(tomorrow), *nakum*(next year).

- *Adverb of Place:* Adverbs of place specify the location of the action. E.g., *Aizawh-ah, ramhnuaiah.*

- *Adjectival Adverb:* It describes the manner in which something is done or occurs and also expresses the feature or quality of the subject. E.g., *Kimi chu a ṭhu hnur mai.* In this sentence, *hnur* conveys that Mawii(Noun) is a big fat lady and she sits improperly.

- *Double Adverb:* Double Adverbs are Adverbs that are used consecutively two times in a sentence to exaggerate the action. E.g., *Chiam chiam, vak vak.*

- *Emphatic Adverb:* Emphatic adverbs are two-word phrases that are used to accentuate the activity in a sentence. E.g., *He pindan chu a va up* ***cherh churh*** *ve* (This room is so stuffy).

  Zarzova and Zikpuia have provided further classifications of adverbs in the Mizo language, including verbal adverbs, nounal adverbs, demonstrative adverbs, adverbs of frequency, and interrogative adverbs.

Conjunction: A conjunction connects words, phrases, or clauses and reveals their relationship. E.g., *chuan, leh*(and)

Interjection: Expresses a spontaneous feeling or reaction. E.g., *Ka rei!, Eheu!*

Postposition: It is equivalent to the English preposition but it comes after the noun or pronoun. So it is called a postposition. E.g., *a, ah, hmain, hnuaiah, chungah,* etc.

3.4 Unique Features of Mizo Language

Across numerous dimensions, Mizo sets itself apart from a plethora of other languages, showcasing a unique and distinct linguistic character, rich with its own set of intricacies and nuances. Below, we delve into the discussion of certain traits, primarily drawing from Mizo Grammar books [120][124][125][128][126][127]:

Mizo language is a tonal language. The meaning of a word could be different depending on the tone of uttering and the context in which the word is being used. According to Fanai[129], tones of the Mizo language can be basically classified into four lexical tones such as low, high, falling, and rising. Some linguists further add sub-divisions viz. short and long for all these basic four classifications, thus differentiating into 8 different tones. The following example shows how the meaning of a single word can change depending on its tone:

Ban - Long High tone - 'arm' - Noun

Ban - Low tone - 'to excommunicate' - Verb

Ban - Falling tone - 'to stretch out' - Verb

Ban - Low tone - 'pillar' - Noun

Sa - Short falling tone - 'meat' - Noun

Sa - Long high tone - 'hot' - Adjective

Sa - Short low tone - 'to build' - Verb

Word order in Mizo language: Generally, Mizo language follows Object–Subject-Verb, as in *Chaw*|Object *ka*|Subject *ei*|Verb, means 'I eat food'. However, several linguists have stated that the Mizo language has "free word order"[129]. That is, the subject, object, and verb in a sentence can be put in any order and yet be meaningful. For instance, consider the following sentences:

Mawiin lehkhabu a chhiar - Subject – Object - Verb

Lehkhabu Mawiin a chhiar - Object – Subject - Verb

A chhiar lehkhabu Mawiin - Verb – Object - Subject

Lehkhabu a chhiar Mawiin - Object – Verb - Subject

All these sentences indicate that 'Mawii reads book'.

The gender of a person can easily be identified. Another distinguishing aspect of Mizo is that the gender of a person can usually be determined by the last letter of that person's name. A person with a name that ends in *'a'* is male, whereas a person with a name that ends in *'i'* is female. For example, *'Lalmawii'* is a female, and *'Lalmawia'* is a male. However, there are a few exceptions that do not adhere to the typical standard Mizo naming conventions.

Agglutinative property: The Mizo language may be classified as an agglutinative language since it follows the process of agglutination [130]. The formation of words can be accomplished by combining different morphemes, with each morpheme preserving its pronunciation and meaning when taken individually. For example:

Lukhum (Cap) : *Lu* (Head), *Khum* (to wear on the head)

Ambiguity due to location: (i) Names for identical objects can vary depending on where they are. For instance, the hair that grows on the head is called *sam,* whereas the hair that grows on other areas of the body is called *hmul.* Therefore, it would be incorrect to say *"Ka chawn sam"* rather, *"ka chawn hmul"* is the correct way to phrase it. (ii)The same word can belong to a different part of speech depending on its location in a sentence. e.g *Chu thil ri chu a ring hle mai.* (That sound is very loud) Consider the word – *'chu'.* In its first occurrence, it is a demonstrative pronoun. In its second occurrence, it is a demonstrative adjective.

Verbs by themselves do not indicate tense. The tense of a verb is indicated by using either an auxiliary verb or an Adverb of time. For examples,

Ka kal(verb) tawh (I went).

Ka kal(verb) mek (I am going).

Ka kal(verb) dawn (I will go).

In the above sentences, the auxiliary verbs *tawh, mek* and *dawn* indicate past, present, and future tense respectively. Consider the following sentences,

Niminah Sikul ka kal. (I went to school yesterday)

Tunah khawnge i awm? (Where are you now?)

Nakkumah lirthei lei ka lei ang. (Next year, I will buy a vehicle)
In the above sentences, adverbs of time *Niminah,Tunah, Nakkumah* indicate past, present, and future tense respectively.

Nouns and pronouns must sometimes be used jointly in some sentences. For instance, "Mawii sleeps" should not be translated as *"Mawii mu."* The proper translation would be *"Mawii a mu."*. In this sentence, *Mawii* is a noun and *'a'* is a pronoun.

Roles of affixes: By inserting a prefix or suffix into the word, certain parts of speech can be created. They are briefly discussed below.

When the prefix 'in' is added to an intransitive verb, it becomes a reflexive verb, reciprocal verb, or co-active verb. For example:
Banah ka intauh (Reflexive Verb).
Thanga leh Sangi an inhau (Reciprocal Verb).
Puitling an insual (Co-active Verb).

The suffix 'a' when added to a Noun, verb, adverb, or adjective turns the word into different kinds of adverbs. For example,
Dawhkâna chemte kha han le teh. (Noun to adverb of place)
Nilainia lo kal a tum. (Noun to adverb of time)
Saisira mut loh tur.(Adjective to adverb of manner)

The suffix 'ah' when added to noun, proper noun, pronoun, and adjective turns the word into different kinds of adverbs. For example,
Thenzawlah in hmun ka lei. (Noun to adverb of place)
Tûkinah ruah a sûr. (Noun to adverb of time)
Liana'n Thangi chu nupuiah a nei. (Noun to specifying adverb)

The suffix 'in' (low tone) is a nominative case marker. When it is added to a noun, pronoun, or adjective, it turns the word into a subject. For example,
Bâwngin hnim a pet.

The suffix 'in'(high tone) when added to a pronoun, adjective or verb turns

the word into a specifying adverb. For example,

Keimahin hna ka thawk.

Zangthalin mu suh.

The suffix 'san' is called an exclusion marker. When it is added to a verb or nounal verb , it turns the word into an 'exclusion verb'. For exampls,

Chaw min eisan.

The suffix 'tir' is called a causative marker. When it is added to an intransitive verb, it turns the word into a transitive verb. For example,

Zirtirtuin naupangho a lâmtir.

The suffix 'ho' when added to a noun, verb, adjective, noun phrase, or noun clause turns them into a plural. For example,

Naupangho an infiam. (*naupang* – child, singular; *naupangho* – children , plural)

The suffix 'tê'(long tone) when added to a noun is called 'small species marker'. It denotes a smaller species of a given object. E.g., *chemtê,artê,vawktê.*

The suffix 'pũi'(long tone) when added to a noun is called big species marker'. It denotes a bigger species of a given object. E.g., *chempũi,arpũi,vawkpũi.*

The suffix 'pùi' (low short tone) when added to a noun or adjective in subjunctive form turns the word into a description of equality or a contemporary. E.g., *chipùi*(same clan), *hnampùi*(same/fellow countryman), *indianpùi*(fellow indian), *mizopùi*(fellow mizo).

The suffix 'siak' is used along with a prefix 'in'. When it is added to a verb after adding a suffix 'in' it turns the verb into a word describing a competition. For example,

zai (sing) : *inzáisiak* (a singing competition)

tlân (run) : *intlânsiak*(a running competition)

The suffix 'na' when added to a verb turn it into a verbal noun that has both

the properties of a noun as well as a verb. For example, *I ke tuamna kha thlâk tawh rawh.* (verb to an instrumental case)

In lênna chu a nuam em? (verb to the Noun of place)

The suffix 'na' also creates an abstract noun from verbs and adjectives. For example,

Hmangaih (Verb) means 'to love' ; *hmangaihna*(love) is a noun

Mizo has its own distinct parts of speech.: Mizo is unique among many languages in that it contains a number of parts of speech that are not common in other languages. Some of them are highlighted here:

-Nouns can also be used as a verb as in Nounal Verb. E.gl,*Ka Aizawl dawn. Aizawl* is a noun but acts as a verb in this sentence.

-In a sentence, nouns can also function as an adjective as in Nounal Adjective. E.g *Thing dawhkan.*
Thing(wood) is a noun, but it plays the role of adjective here.

-Double Verb is another distinct part of speech in the Mizo language; it is a verb written twice consecutively to indicate repeated actions. -Adjectival verbs: Some adjectives can be used as a verb in a sentence. e.g *Liana uitê chu a fing hle.* (*fing*(adj)- which means clever here is used as a verb) -Verbal adverb : Many verbs are used as adverb without any change of root word form. E.g., *Bungrua ka thiar chhuak.* In this sentence, chhuak(verb) which means 'to go out' is used as an adverb.

-Numeral adjectives are repeated to indicate an emphasis. For example, *pakhat*(one), *pahnih*(two), *pathum*(three), etc., can be repeated as in *pakhat khat, pahnih hnih, pathum thum.*

-Use of double adjectives to make plural form : When some adjectives are repeated, it means more than one objects in the sentence. E.g., *Artui lian lian kha ei rawh.*(Eat the eggs, the big ones). In this sentence, *lian*(big,adj) is repeated to indicate more than one eggs.

Compound noun: Compound nouns are nouns which are formed by joining more than one words which can be solid (without a space in between the noun words), spaced(with a spaced in between) or hyphenated(with a space in between).For example,
kalkawng(kal + kawng) - a solid compound noun.
biak in (biak + in) - a spaced compound noun.
Puan- ṭhui-khawl - a hyphenated compound noun.

3.5 Challenges in Mizo language from Computational Point of view

The Mizo language *(Mizo ṭawng)* is both tonal and morphologically intricate, making language processing tasks in Mizo a significant challenge. In this section, we emphasize some of the difficulties encountered when undertaking tagging tasks in the Mizo language.

Ambiguity is always the main major challenge in tagging works; the same is true for the Mizo language. A word with the same spelling and same pronunciation may indicate different meanings depending on the context of the sentence. For example, a Mizo word *'Pu'* may mean a grandfather as in *Ka pu hming chu John a ni*, which means 'My grandfather's name is John'. The same word *'Pu'* may also mean 'to carry on the shoulder' as in *Thing ka pu* i.e 'I carry a log.' In fact, a complete sentence could also be ambiguous. E.g., *Hei hi ka lei a ni.* This sentence could mean 'This is my bridge' or 'This is what I bought'. So, tagging for *'lei'* could be a noun' or 'Nounal Verb'.

It might be challenging to select the right tone for similar words with various meanings in Mizo because it is a tonal language. Moreover, the Mizo language lacks standard or accepted diacritical indicators to distinguish between different tones. The tone marker '^'(caret or circumflex symbol) is the only symbol that is

accepted in general to denote all the long tones in the language. This tone marker, however, is not always precise and can occasionally cause confusion because it is applied to all four levels of tone previously discussed. This complicates Mizo's language processing even further.

A different concern arises from the fact that Mizo writers have varying viewpoints, leading to a lack of consistent writing style. While some authors may use the phrase *'hrilh fiah'*, others might opt for the combined form *'hrilhfiah'*. Even language experts have differing viewpoints, making it challenging to develop solid and concrete policies and guidelines for writing words.

Another major impediment to language processing in Mizo is the lack of tagged corpora. There are not many resources available for the Mizo language. To the best of our knowledge there in no publicly available tagged corpus that is readily available to the public. A standard corpus for different domains needs to be created which is universally accepted by experts in the field.

Mizo grammar intricacies can occasionally become quite intricate. Various grammar experts hold differing opinions regarding the usage of specific words and structures. This complexity hinders the creation of a comprehensive grammar framework when it is required.

In conclusion, this chapter has delved into a comprehensive exploration of the Mizo language, encompassing its distinctive linguistic characteristics, unique features, and computational challenges. This chapter serves as a pivotal foundation for the subsequent analysis of Part-of-Speech (POS) tagging for the Mizo language. By thoroughly understanding the intricacies of Mizo's syntax, morphology, and semantics, we are better equipped to design, develop, and fine-tune effective POS tagging models tailored specifically to the nuances of this language. The challenges identified in this chapter, ranging from scarce linguistic resources to intricate agglutinative structures, underscore the need for innovative and context-sensitive approaches to POS tagging in the Mizo language. This groundwork not only facilitates the enhancement of natural language processing tools for Mizo

but also contributes to the broader advancement of computational linguistics, as solutions developed here may find applications in other morphologically rich languages with similar characteristics. Therefore, this plays a pivotal role in guiding the subsequent stages of this research, fostering the development of accurate and contextually sensitive POS tagging methods that can significantly impact the analysis, understanding, and generation of text in the Mizo language.

Chapter 4

TAGGING WITH HIDDEN MARKOV MODEL

4.1 Introduction

In the past, a popular technique for assigning part-of-speech (POS) tags to words relied on rule-based tagging, which entailed the utilization of manually crafted rules to assign tags to words. However, this approach can be time-consuming and prone to errors, especially for languages with complex grammar and syntax.

To address these issues, a stochastic-based POS tagger called the Hidden Markov Model-based tagger, abbreviated as the HMM-based tagger, has been developed. Stochastic taggers are supervised models that assign tags to new words based on the frequency or probability of tags found in the training corpus. These taggers employ statistical techniques to learn patterns and relationships between words and their corresponding POS tags in the training data. HMM-based taggers offer greater flexibility and adaptability compared to rule-based taggers, enabling them to handle diverse languages and text genres effectively. They have been proven to be effective in various NLP tasks and are widely utilized in both academic research and commercial applications.

The concept of the Hidden Markov model (HMM) was first introduced by Baum and Petri [131]. It was named after the Russian mathematician Andrey Andreyevich Markov, who pioneered most of the related statistical theory. Hidden Markov Model (HMM) is one of the prevalent statistical or probabilistic-based models. This model does not necessitate a substantial amount of expert knowledge concerning the morphological structure of the language. HMMs have a wide range of applications, including machine translation, speech analysis, handwritten recognition, signal processing, sequence classification, time series analysis,

POS tagging, phrase chunking etc. They also found applications in the area of bioinformatics, human activity recognition, musicology, and network analysis[132]

In this chapter, our primary focus will be on the application of Hidden Markov Models (HMM) for tagging the Mizo language. We will explore the fundamental principles that underlie the use of HMMs in this context. Additionally, we will provide a comparative analysis between the unigram tagger and the bigram HMM tagger, shedding light on their respective strengths and weaknesses.

4.2 Related works

In this section, we present the research works related to a part-of-speech tagging system for different languages. Since the inception of coding of part-of-speech tagging systems in the year the 1960s, lots of improvements have been made with different techniques and methods. A system based on the statistical method is one of the most popular tagging systems for different languages. HMM-based part of the speech tagging method was introduced[38] in the mid-1980s

HMM-based Part of speech tagger for Arabic is discussed in [133]. In that paper characteristics of Arabic languages and 55 tagsets have been proposed. They have developed a 9.15 MB corpus of native Arabic articles. Words of 23554 verbs, 27594 nouns, 5384 proper nouns, and 5722 adjectives were chosen to train the tagger. 944 words were used as tested corpus and achieved an accuracy of 97%. Statistical POS tagging system in Persian text is presented in [134]. They have created a tagged corpus and evaluated statistically based TnT tagger on the Persian language. The experiments were repeated several times on 80% and 15% of the corpus as training data and test data respectively and the data were selected randomly. They obtained an overall average accuracy of 96.59%.

A part-Of-Speech tagging system for the Urdu Language based on a statistical model is discussed in [135]. In this paper, supervised learning, the N-gram Markov model tagging method was used. The experiments were performed based on the

unigram, bigram, and backoff methods on small and large tagset. They achieved higher accuracy with a smaller tagset and with the backoff method, they could achieve 95% accuracy. Arabic Part of Speech Tagging based on Parallel HMM is discussed in [136]. In this paper, they proposed a new approach tagging system that relies on two HMM working together in parallel in the system. The first one is the main model and the second model is used as a reference for low probabilities tags. Both models are trained using the dual corpus. To overcome the time complexity, the system was implemented using a multithreading approach. The average accuracy of 75.38% was obtained on a small dataset. Though the concept is novel, the system is tested on a very small dataset that consists of 40 numbers of sentences (845 words). The performance of the system is yet to be trialed on the large dataset to see the actual performance of this method.

Denis et. al [137] discuss the presented POS tagger for the Indonesian language based on the HMM N-gram (bigram and trigram) approach and Viterbi algorithm. They have compared HMM bigram and HMM trigram on the Indonesian language corpus and found that HMM bigram scored better with an accuracy of 77.56% whereas 61.67% accuracy was obtained with HMM trigram. N. Joshi [104] presented a POS tagging system for the Hindi language based on Hidden Markov Model. 15,200 sentences were utilized to train the system and the IL POS tag set was used in the system. They obtained an accuracy of 92% on the test data.

HMM-based POS Tagging system for Kayah Language is discussed in [138]. They have developed 16 tagsets to disambiguate words in the Kayah language and they achieved an average accuracy of 87%. J. Singh et. al[139] presented part-of-speech tagging in Marathi using statistical method. They have implemented and compared unigram, bigram, trigram, and basic HMM. They used a tagset developed by IIIT Hyderabad, and a test corpus of 25744 words (1000 sentences) was developed to see the performance of the system. They achieve accuracy of 77.38%, 90.30%, 91.46%, and 93.82% for unigram tagger, bigram tagger, trigram tagger, and HMM-based tagger, respectively.

Mohammed[140] discussed a stochastic-based POS tagging system for the Somali language, which is a low-resource language. He presented the first POS tagging system for Somali using different approaches such as HMM, Conditional Random Fields, and Neural Networks. 14369 words were used to train the system and obtained an average accuracy of 87.51%. There are many more papers related to part-of-speech tagging based on Hidden Markov Model(HMM) for different languages[38][141][142][143]

4.3 Building the Resources

The development of a dataset and tagset is a crucial aspect in building language resources for natural language processing tasks. In this section, we will provide a brief overview of the process involved in defining the tagset, creating the dataset and present some statistics related to these resources.

4.3.1 Formulating the Tagset

Tagset is a list of collections of tags or labels, designed to indicate the morphological classes of each word in the sentences. Tagset are usually designed for specific languages since the morphological structures of languages are different for different languages. It is essential to design a proper tagset to signpost the grammatical information about each token in the corpus.

To create a tagset for POS tagging in the Mizo language, it is important to carefully analyze the language and identify the various grammatical categories that exist. This can be done by analyzing the syntax and morphology of the Mizo language, as well as looking at examples of real-world text. The tagset should be comprehensive enough to cover all the possible parts of speech and their various inflections and variations.

By creating a comprehensive and accurate tagset for POS tagging in the Mizo

language, it will be possible to develop more effective and efficient natural language processing applications that can process and analyse text written in the Mizo language with greater accuracy and precision.

For this research work, in order to increase efficiency and accuracy, we have combined some of the fine-grained tags. For example, material nouns, countable nouns, common nouns, and concrete nouns are classified under the category of Common Noun. Following extensive research and analysis of the Mizo language's morphological structure, tagsets consisting of 45 tags covering all the grammatical information of the tokens in the corpus have been proposed, as shown in Table 4.1.

4.3.2 Collecting the Mizo Text

Corpus, in the computational linguistics context, is a collection of structured text data. They are usually designed for a specific purpose with a specific format. It is an essential task to build a large annotated corpus to perform part of speech tagging using training-based techniques. They are the main language resources and knowledge beds for language processing to perform statistical analysis.

To the best of our knowledge, there is no proper POS tagged corpus in the Mizo language till today. So, the raw digital texts are collected from different sources and from different topics, the majority from the Vanglaini daily news and articles (online version). This collection of texts includes different topics such as daily news, health, culture, politics, sports, etc.

To prepare raw texts in the Mizo language for computational processing, it is necessary to clean and normalize them properly. Therefore, during the process of compiling the text, unnecessary punctuation marks are eliminated from the sentences, spelling errors are rectified, and writing styles are standardized.

Table 4.1: List of proposed Mizo tagset

Tags	**Descriptions**	**Examples**
PPN	Proper Noun	*Sanga* (person's name)
CMN	Common Noun	*Pa* (father), *Thei* (fruits)
ABN	Abstract Noun	*Beiseina* (hope), *Hmangaihna* (love)
PSP	Personal Pronoun	*An* (They), *Kan* (We), *Ka* (I), *I* (You)
POP	Possessive Pronoun	*An* (Their), *Kan* (Our), *Ka* (My), *In* (Your)
RLP	Relative Pronoun	*Hi, Chu, Saw, Kha, Khi, Khu*
IP	Interrogative Pronoun	*Eng* (what), *Tu* (who)
MP	Demonstrative Pronoun	*Hei hi* (This is), *Saw saw* (That is)
JJ	Adjective base form	*Te* (smal), *Tawi* (short), *Thlum* (sweet)
MJJ	Demonstrative Adjective	*Khung khu, Heng hi, Khang kha*
DJJ	Double Adjective	*Mawi mawi* (big ones), *Te te* (small ones)
IJJ	Interrogative Adjective	*Tu* (who), *Eng* (what)
CJJ	Comparative Adjective	*Sang zawk* (taller), *Chak zawk* (stronger)
SJJ	Superlative Adjective	*Sang ber* (tallest), *Chak ber* (strongest)
VB	Verb base form	*Ziak* (write), *Ei* (eat)
NVB	Nounal Verb	*Aizawl*, Eg:- *Ka Aizawl dawn* (I am going to Aizawl)
IRB	Interrogative adverb	*Engtinnge*
DVB	Double Verb	*Kal kal* (go oftenly)
RB	Adverb base form	*Chak* (fast), Eg:- *A kal chak hle* (He goes very fast)
DRB	Double Adverb	*Vak vak*, Eg:- *A au vak vak* (He is shouting repeatedly)
MRB	Demonstrative Adverb	*Lo, han.* Eg:- *Ka han kal ang* (I am going there)
PPT	Postposition	*Atan, vang*
CC	Coordinating Conjunction	*Leh* (and), *emaw* (or)
UH	Interjection	*Awi! Karei!*
SPRB	Specifying Adverb	Hmanga, kaltlanga
PT	Particles	*Te, a, ni.* Eg:- *Miza naupang fel tak a ni.*(He is a good boy)
SYM	Symbol	$, #,, &, %.
,	Comma	,
.	Fullstop	.
:	Colon	:
;	Semicolon	;
?	Question mark	?

QM	Quotation Mark	“ ”, ‘ ’
CD	Cardinal number	4,6,7, *pali* (four), *pangana* (fifth), *sawm*(ten)
NG	Negation	*Lo* Eg:- *Mawi lo* (not beautiful)
ET	Date	8th August, 2021
RBP	Adverb of Place	*Champhaiah* (at Champhai), *dawhkan chungah* (on the table)
RBT	Adverb of Time	*Zingah*(in the morning), *Chawlh-niah*(on sunday)
SF	Suffix	*Ah, te, in*
AT	Article	*Chu, chuan*
RBM	Adverb of Manner	*Zangthal*(to lie on the back)
FW	Foreign Word	Conference, General, Camping
CRB	Comparative Adverb	*Tlan chak zawk* (run faster)
SRB	Superlative Adverb	*Tlan chak ber* (run fastest)
VBN	Verbal Noun	*Kalna* (means of going)

4.3.3 Tokenization

Tokenization involves breaking down the original text into smaller units called tokens, which are more manageable for processing. During tokenization, words within sentences are divided and separated by a single space. In addition to regular words, punctuation marks like commas, periods, colons, and semi-colons are also tokenized. By tokenizing these punctuation marks, they are distinguished from the words in a sentence. This separation aids the part-of-speech (POS) tagger in correctly determining the grammatical function of each word in the sentence.

4.3.4 Manual Tagging

In order to address the absence of a publicly accessible labeled Mizo corpus, a new corpus is created through meticulous manual tagging. This process involves painstakingly annotating each word or token in the tokenized text with its appropriate part of speech. Each word is manually tagged with its appropriate part-of-speech (POS) tag, with a slash(/) put between the word and its corresponding tag. Each sentence is completed with a period(.). This manual tagging

process ensures that each word in the corpus is accurately assigned its respective part of speech, allowing for more precise linguistic analysis and processing. An illustrative example of a tagged sentence within the corpus is provided below:

Liana/PPN chu/AT naupang/CMN fel/JJ tak/RB a/PSP ni/VB ./.

The manual task of annotation was done carefully with the 45 tagset to handle different ambiguities that exist in the Mizo language and a lot of help from linguistics experts was obtained to establish a reliable corpus for Mizo POS tagging. In this research work, a corpus consisting of 23,319 words (688 sentences) has been built. A summary of the developed Mizo corpus during this research work is given in Table 4.2

Table 4.2: Summary of the developed Corpus

Particulars	**Count**
Total number of words (including symbols)	23,319
Total no. of sentences	688
No. of unique tags	45
No. of unique vocabulary	4,442

4.4 Development of Hidden Markov Model-based POS Tagging System

Any tagging model that contains frequency or probability in some way can be properly classified as stochastic. i.e it uses frequency, probability, or statistics to assign a tag to the term. The primary goal of this chapter is to develop a stochastic-based model known as bigram Hidden Markov Model-based tagger.

When given a sequence of units (words, letters, morphemes, etc.), a Hidden Markov Model calculates a probability distribution over various label sequences and selects the optimum label sequence [144]. However, in order to comprehend Hidden Markov Model (HMM), it is necessary to be familiar with Markov Chains.

4.4.1 Markov Chain

Markov chains serve as the foundation for the Hidden Markov Model. Markov model is a stochastic model that is used to simulate events or states that change randomly. A Markov chain makes the assumption that in order to predict the future state in the series, only the current state is important [145]. This means that the only thing that matters for the future is the present. Every state that has come before the current state has absolutely no bearing on what will happen in the future.

Let x1, x2, and x3 represent the past, present, and future events at times t-1, t, and t+1, respectively. According to the Markov model, x3 is dependent only on x2, and x1 does have any effect on x2. It does not matter how the current state x2 has arrived. The graphical representation is shown in Fig. 4.1 For instance, a

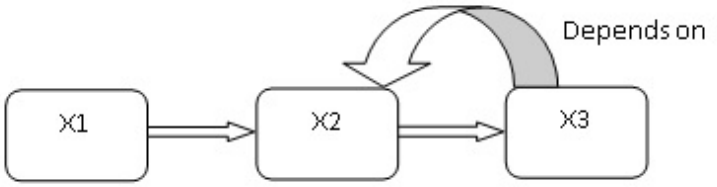

Figure 4.1: Graphical representation of Markov Chain

weather condition may be any of the three potential states, such as sunny, rainy, or cloudy. The Markov Chain model would only look at today's weather in order to anticipate tomorrow's weather, as yesterday's weather has no bearing on the forecast.

The trait of memorylessness is the defining feature of a Markov Chain model. The model embodies the Markov assumption regarding the probability of this sequence, which states that for predicting the future, only the present is relevant. More formally, consider a sequence of state variables $X_1, X_2, X_3, ..., X_i$, the probability of X_i is given as

$P(X_i|X_{i-1}, X_{i-2}) = P(X_i|X_{i-1})$

A Markov Chain model is defined by the following elements:

i) A set of N states.

ii) *Transition Probability matrix:* P is a transition probability matrix in which each P_{ij} reflects the likelihood of transitioning from state i to state j, such that $\Sigma_{j=1}^{n} = 1$

ii) *Initial probability distribution:* It is the likelihood that a specific state will be the Markov chain's initial condition. Some j states may have $\pi_j = 0$, indicating that they are not initial states.

The example that follows, which depicts the states and transition probabilities in Fig. 4.2, may help to further explain the notion of a Markov model. Cloudy, Raining, and Sunny, are the states and the decimal numbers represent the transition probabilities between the states. For instance, if it is cloudy today, there is a 0.5 probability that tomorrow will be sunny.

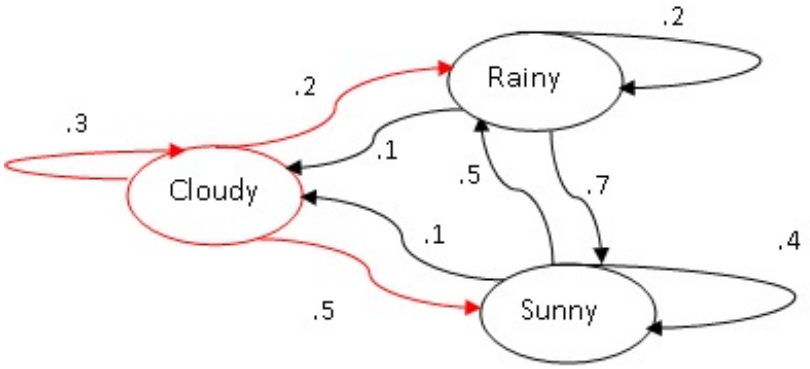

Figure 4.2: A Markov chain with states and transitions

4.4.2 Hidden Markov Model

Sometimes, it is required to predict a sequence of states that are not directly observable in the environment, i.e., the states are hidden. Instead of hidden states, we are given evidence or observable events. Depending upon the evidence or observable events, we have to find out the hidden states.

In the Fig. 4.3 x_1, x_2, x_2, are hidden states and observable states are y1, y2,

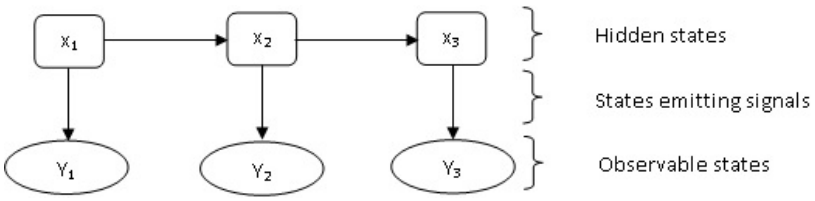

Figure 4.3: A pictorial representation of an HMM

and y3. Though we are given another sequence of states that are observable in the environment, these hidden states have some dependence on the observable states. In other words, it is necessary to find the corresponding hidden states from the visible states that the hidden states emit.

To give an alternative explanation, let's once more consider various weather conditions such as cold, sunny, and cloudy. Suppose, these weather conditions are not directly observable (i.e., it is not possible to say weather conditions directly, since it is hidden). Instead, we can observe someone's attire, such as a person wearing a coat, carrying an umbrella, or wearing shorts to infer the weather conditions. Therefore, we need to deduce the hidden states(the weather conditions), based on these observable states.

In other words, a hidden Markov model is a type of statistical model for depicting probability distributions over a sequence of observation events [146]. It is a model that assumes the system being described is a Markov process with unobserved (hidden) states. The basic architecture of the HMM is shown in

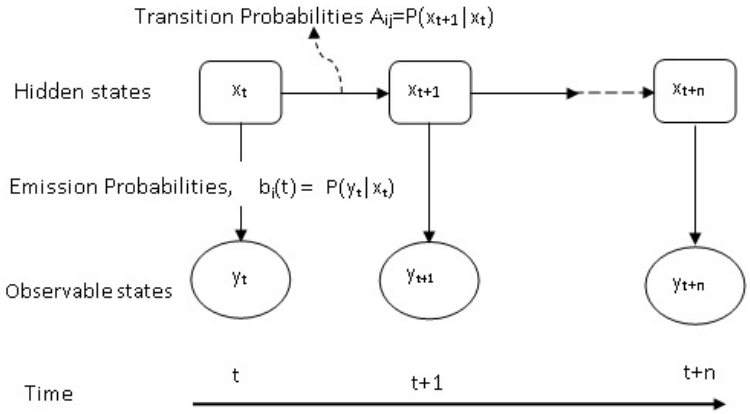

Figure 4.4: Basic architecture of HMM

Figure. 4.4 Hidden Markov Model has the following components:

$X = x_t, x_{t+1}, ..., x_{t+n}$	: A set of N states (hidden states)
$A = a_{11}, a_{12}, ..., a_{ij}, ..., a_{nn}$	: A is a transition matrix in which each a_{ij} reflects the likelihood of transitioning from state i to state j, such that $\sum_{j=1}^{n} = 1$
$Y = y_t, y_{t+1}, ..., y_{t+n}$	: A sequence of observable symbols at each time interval
$B = bj(t)$	: *Emission probabilities:* A sequence of observation likelihoods, each expressing the probability of an observation y_t from the state j. This can be expressed as $P(y_t\|x_t)$
$\pi_1, \pi_2, ..., \pi_n$	: Initial probability distribution: It is the likelihood that a specific state will be the Markov chain's initial condition. Some j states may have $\pi_j = 0$, indicating that they are not initial states.

4.5 POS tagging with HMM

The observation sequences and a set of possible states must both be available in order to model any problem using a hidden Markov model. The states in an HMM are hidden. In the part-of-speech tagging problem, the observations are the words themselves in the given sequence. As for the states, which are hidden, these would be the POS tags for the words.

In the POS tagging problem, we have the following values:

Q: Set of possible Tags

A: The A matrix contains the tag transition probabilities $P(t_i|t_{i-1})$ which represent Transition Probabilities from one tag t_{i-1} to another t_i. For example, $P(Adjective|Pronoun)$ is the probability that the current tag is an adjective, given the previous tag is a Pronoun. This can be calculated as:

$$P(Adjective|Pronoun) = Count(PronounandAdjective)/Count(Pronoun)$$

O: Sequence of observation (words in the sentence)

B: The B emission probabilities, $P(w_i|t_i)$, represent the probability that the word is w_i (say eating), given a tag t_i(say Verb), The emission probability B[Verb][eating]

is calculated using:

$P(eating|Verb)$: $Count(eating\&Verb)/Count(Verb)$.

It must be noted that we get all these Count() from the corpus itself used for training.

Initial Probability: Probability of the initial word

4.5.1 The Maths

The use of HMM for tagging system is a special case for Bayesian inference[1], a paradigm that is trying to choose the best tag sequence that corresponds to the sequence of words in a corpus. It is a task of finding the sequence of POS tags t_1^n that is the most probable tag sequence from a given word sequence w_1^n. So, we have,

$$t_1^n = argmax_{t_1^n} P(t_1^n|w_1^n) \tag{4.1}$$

Here, t_1^n is a sequence of tag $(t_1...t_n)$ and w_1^n is a sequence of words $(w_1...w_n)$.

A conditional probability, using Bayes' rule, is given as follows:

$P(a|b) = P(b|a)P(a)/P(b)$

So, By using Baye's rule for conditional probability, Equation 3.1 can be expressed as follows:

$$t_1^n = argmax_{t_1^n} P(w_1^n|t_1^n)P(t_1^n)/P(w_1^n) \tag{4.2}$$

For each tag sequence, $P(w_1^n)$ remains the same, so it can be neglected. Therefore,

$$t_1^n = argmax_{t_1^n} P(w_1^n|t_1^n)P(t_1^n) \tag{4.3}$$

Here, $P(w_1^n|t_1^n)$ is referred to as the likelihood of the word string, and $P(t_1^n)$ is

called the prior probability of the tag sequence. The HMM is predicated on two assumptions: The first assumption states that the word's probability of occurring is solely determined by its own tag. So,

$$P(w_1^n|t_1^n) \approx \prod_{i=1}^{n} P(w_i|t_i) \tag{4.4}$$

The second supposition states that the tags' probability of occurring is determined by the previously fixed n number of tags. For this research work, the bigram model has been considered, in which the probability of a tag is determined solely by the preceding tag. So,

$$P(t_1^n) \approx \prod_{i-1}^{n} P(t_i|t_{i-1}) \tag{4.5}$$

The following equation is derived by substituting Equation 3.4 and Equation. 3.5 into Equation 3.3 :

$$t_1^n = argmax_{t_1^n} P(t_1^n|w_1^n) \approx argmax_{t_1^n} \prod_{i=1}^{n} P(w_i|t_i)P(t_1|i_{i-1}) \tag{4.6}$$

The bigram tagger employs Equation 3.6 to identify the most likely tag sequence. It contains two kinds of probabilities, P($w_i|t_i$), i.e., emission probabilities, and P($t_i|t_{i-1}$), i.e., tag transition probabilities.

Transition probability: The probability of a particular tag occurring in a sequence, given the preceding tag, is referred to as Transition probability. The transition probabilities can be calculated using the following equation:

$$P(t_i|t_{i-1}) = C(t_{i-1}, t_i)/C(t_{i-1}) \tag{4.7}$$

Here, C(t_{i-1}, t_i) denotes the number of times the current tag occurs alongside the previous tag in the training corpus, and C(t_{i-1}) is the previous tag's frequency count in the corpus.

Emission Probability: The emission probability determines the most appro-

priate tag for the specific word based on the number of occurrences of the word. These can be calculated using the following equation:

$$P(w_i|t_i) = C(t_i, w_i)/C(t_i) \tag{4.8}$$

$C(t_i, w_i)$ denotes the number of times the current tag is associated with the present word. $C(t_i)$ denotes the present tag's frequency.

The objective of the Hidden Markov Model is to find the highest probable tag sequence. A simple brute force method to solving this problem would be prohibitively costly. For instance, consider the following sentence:

"I love you,"

Suppose, we have two possible tags for these three words: PP and VB. Some of the probable tag sequences would be, {PP, PP, PP}, {PP, PP, VB}, {PP, VB, PP}, {PP,VB, VB}, {VB, PP, PP}, etc. There would be 23=8 potential sequences.

This may not appear to be a massive chunk, but the number of sequences will expand exponentially as the number of observations increases over time. The total number of possible tag sequences for a sentence generated by HMM would be t^n, where t is the number of probable tags, and n is the number of observations(words).

Due to the exponential growth of the number of sequences, it is not feasible to use a brute force strategy for sentences of reasonable length, as it would take an unreasonable amount of time to execute. Therefore, as a substitute for this brute force approach, the Viterbi Algorithm has been utilized to efficiently determine the most probable tag sequence.

4.5.2 The Viterbi Algorithm

In part-of-speech tagging systems, the Viterbi algorithm is the most commonly used decoding algorithm for Hidden Markov Models. This dynamic programming algorithm is employed to ascertain the most probable sequence of hidden states, which is also referred to as the Viterbi path. The working of the Viterbi algorithm is described as follows:

i) Setting up of a probability matrix, wherein all observations o_t are listed in a single column, and each state q_i is listed in a separate row.

ii) Given the A and B probability matrices, a cell in the matrix $v_t(j)$ indicates the likelihood of being in state j after t observations and traversing through the sequence with the highest probability.

iii) The following equation is used to figure out the value of each cell:

$$v_t(j) = \max_{q_1 \ldots q_{t-n}} P(q_1 \ldots q_{t-1}, o_1, o_2 \ldots o_t, q_t = j | (A, B)) \tag{4.9}$$

An example of a Viterbi matrix for a sequence of words and the probable tags(states) for each word is shown in Figure 4.5 The tags with the highest probability for each word of the given sequence is shown with the highlighted arrows. A greyed state denotes a probability of zero for a given word sequence based on the B matrix of emission probabilities.

The process of calculating each cell value in the Viterbi algorithm is done through iteration. The Viterbi probability, denoted by $v_t(j)$, can be computed as follows for any given state q_j at any given time t:

$$v_t(j) = max_{i=1}^{n} v_{t-1}(i) a_{ij} b_j(o_t) \tag{4.10}$$

So, the Viterbi probability is calculated by multiplying three components: (i) Probability of the prior Viterbi path $v_{t-1}(i)$ ii) The transition probability from the

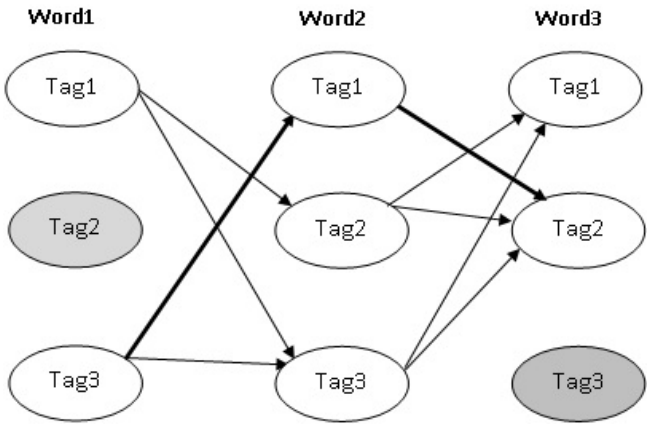

Figure 4.5: Viterbi matrix with possible tags for each word

prior state to the present state, a_{ij}. iii) The probability of the state observation symbol o_t, given the present state j, $b_j(o_t)$

4.6 Experiment Set Up and Result Analysis

This section discusses briefly the experimental works performed based on the manually annotated corpus consisting of 23,319 words and presents the results and their analysis.

We have also examined the performance of the Unigram tagger (1-gram tagger), one of the simplest stochastic-based POS taggers, for comparison purposes. A unigram tagger takes a single word as context to determine the POS tag. It identifies the most probable tag for each word in a training corpus and then assigns tags to new tokens based on this information. The Unigram tagger is a basic and straightforward approach to part-of-speech tagging but is limited in its ability to handle complex linguistic phenomena and contexts. More advanced taggers are often preferred for tasks that require higher accuracy and contextual understanding.

4.6.1 Tagset Distribution in the Corpus

Out of 45 tagsets used for labeling words in the corpus, Figure 4.6 presents the tag frequency distribution in the whole corpus with higher than 2% in our complete corpus. It is observed that Verb(VB) has the highest occurrence, followed by an adverb(RB), common noun(CMN) and personal pronoun(PSP).

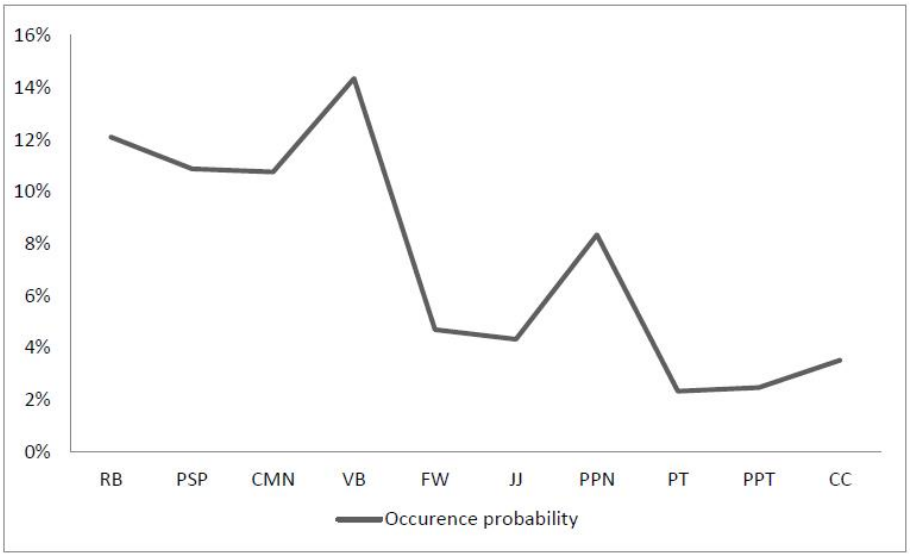

Figure 4.6: Tagset occurrence in the corpus

4.6.2 Transition Probabilities

The transition probability is one of the very important calculations in the bigram HMM tagger. It is the likelihood of the particular tag sequence, considering two tags, i.e tag-tag pairs. As per the Bigram HMM tagger assumption, the probability of a tag depends on the preceding tag. Figure 4.7 depicts the top higher transition probability generated when the splitting ratio is 0.9: 0.1 (only some portion is shown here due to space limitation). It can be seen that the personal pronoun (PSP) following the verb has the highest transition probability.

Figure 4.7: Transition Probabilities

4.6.3 Accuracy of the Taggers

To assess the effectiveness of the two taggers, the corpus that has been manually tagged is divided into two separate sets: one is designated as the training dataset, and the other as the test set. The accuracy of the taggers is evaluated on different sizes of the training set and test set by splitting the corpus into percentage ratios of 70:30,75:25,80:20,85:15,90:10 as training set and test set respectively. The accuracy of the taggers is calculated using the following formula :

$$Accuracy = (No\,of\ correct\,tags/No\,of\ words)X100$$

Table 4.3: Accuracies of the taggers

Train set : Test set	**Unigram**	**Bigram HMM**
70:30	68.61%	72.39%
75:25	69.09%	72.52%
80:20	69.81%	73.98%
85:15	70.53%	74.30%
90:10	70.61%	75.19%

The performance of the unigram shown in the table above is relatively good but sometimes this approach may give tag sequences that conflict with the grammar rules of a language. The unigram tagger points out the most commonly used tag for a particular word in the annotated training data and it utilizes this information to label the word in the unannotated text. For example, if the training set contains a Mizo word in (drink) 20 times tagged as 'verb', and in (house) 10 times tagged as 'Noun', then all the words 'in' in the test set will be tagged as 'verb'. Unlike the Unigram tagger, the bigram HMM tagger depends not only on the frequency of a single word, instead, it considers the probability of a tag sequence with the previous tag. The result obtained is presumed to be more reliable than the unigram tagger. The accuracy increases as the size of the training data increases and in our experiment, the maximum accuracy of 75.19% is obtained when the training data is highest. So, it is expected to improve the performance of the taggers with the increase in the size of the corpus.

4.7 Conclusion and Future Works

It is challenging and exciting to work on the development of language processing tools for under-resource language. In this chapter, we have designed a model for a stochastics-based part-of-speech tagging system for the under-resourced language, in the case of the Mizo language. Preprocessing and cleaning of raw texts collected from different domains have been carried out carefully. A reliable Mizo corpus consisting of 23,319 words was created and annotated the corpus manually with the proposed 45 tags. The performances of the developed taggers were evaluated on this corpus. The experiment was repeated and evaluated with different splitting ratios of the corpus. The experiment results show that the average accuracy of the unigram tagger is 70.61% and the accuracy of the bigram HMM tagger is 75.19%.

There are heaps of space for development in the computational linguistics

field in the Mizo language. The primary drawback affecting the performance of the taggers is the scarcity of available resources. To enhance the accuracy of the tagging system, it is anticipated that increasing the amount of data in the corpus would be beneficial. Additionally, it would be advantageous to endeavor to establish uniformity in writing styles within the Mizo language. It is also possible to use certain rule-based components to detect and correct current model defects. Furthermore, a comparative investigation into the effectiveness of various techniques, particularly statistical approaches, could be pursued in upcoming studies.

Chapter 5

TAGGING WITH CONDITIONAL RANDOM FIELDS

5.1 Introduction

In Chapter 4, we have introduced Hidden Markov Model (HMM)-based POS taggers, which rely on a probabilistic model to assign the most likely POS tags to words in a sentence. However, HMMs have limitations when it comes to capturing the dependencies between neighboring words and the contextual information necessary for accurate POS tagging. Conditional Random Fields (CRFs) offer an alternative approach that addresses these limitations.

HMMs are generative models that join the probability distribution of the observable and underlying hidden state sequences. The observable sequence typically represents the sequence of observations or input data, while the hidden state sequence represents the sequence of latent or unobserved states. HMMs assume that the hidden state sequence follows a Markov process, meaning that the probability of transitioning from one state to another depends only on the current state, not on any earlier states [147]. This property is known as the Markov assumption.

In contrast to Hidden Markov Models (HMMs), Conditional Random Fields (CRFs) are discriminative models that directly estimate the conditional probability of the hidden state sequence based on the observable sequence [148]. CRF algorithms learn the conditional probabilities of each label by considering the preceding labels and observations in the sequence, utilizing a training set comprising annotated sentences. Subsequently, these probabilities are employed to estimate

the most probable label sequence for each word in a given phrase or sentence that was not previously encountered [148].

By leveraging the dependencies between hidden states and observed data, CRFs offer a more flexible and expressive modeling approach. The transition from HMM-based tagging to CRF-based tagging opens up exciting possibilities for improving the accuracy and performance of our tagging systems. CRFs allow us to capture more complex dependencies between tags, as they are not limited by the Markov assumption that HMMs rely on. This flexibility enables us to incorporate richer contextual information and exploit more fine-grained patterns in the data.

In the realm of natural language processing (NLP), where sequence labeling is a fundamental task, CRFs have proven to be a powerful tool. They have been successfully applied to various NLP tasks such as part-of-speech tagging, named entity recognition, syntactic parsing, and semantic role labeling [58]. CRF models may also add characteristics like word embedding and contextual information to increase their accuracy on several NLP tasks [149]. The use of CRFs in these applications has often led to improved accuracy and better generalization.

This chapter offers a detailed exploration of Conditional Random Fields (CRFs) in the context of part-of-speech (POS) tagging specifically tailored for the Mizo language, delving into their concepts and principles. It will also provide a comprehensive examination of the resource development process for training and evaluation. The chapter will further present a thorough overview of the experimental setup, encompassing diverse configurations, parameter selections, and feature engineering techniques utilized for training the CRF models. The obtained results will be meticulously analyzed, with a specific emphasis on performance metrics like accuracy, precision, recall, and F1 score.

5.2 Related works

Many POS taggers based on Conditional Random Field(CRF) have previously been developed for a variety of Indian languages, including Hindi, Marathi, Bengali, Punjabi, Kannada, Tamil, Malayalam, Khasi, Manipuri, and many others. This section highlights a few of them.

Ekbal et al. [150] presented a method for Bengali Part of Speech (POS) tagging using statistical Conditional Random Fields (CRFs). They developed a POS tagger that achieved high accuracy by incorporating a contextual window within the statistical CRF framework. To address unknown word challenges, the researchers implemented various techniques, resulting in a significant improvement in system accuracy. The achieved accuracy for the system was reported to be 90.3%.

In their work, Patel et al. (2008) discussed the machine-learning approach used for Gujarati POS tagging. They employed a CRF model and integrated language-specific rules as features, resulting in an impressive accuracy of 92%. They anticipate even better results with an expanded training dataset.

Pandian et al. [151] presented Language Models for Tamil, focusing on part-of-speech (POS) tagging and chunking, using CRFs. The models were enriched with morphological data, and the authors proposed the integration of an error-correction module like Transformation-based Learning to improve the task. By incorporating "boundary detection before the chunking process," the model could potentially achieve enhanced results. The reported accuracy for their approach was 84.25%.

Nongmeikapam et al. [152] presented a modified feature selection in Conditional Random Field (CRF) based Manipuri Part of Speech (POS) tagging. This study aimed to enhance the efficiency of previous work [153] by experimenting with multiple new features. Additionally, the study incorporated Reduplicated Multiword Expression (RMWE) as another feature, considering that the Ma-

nipuri language is rich in RMWE, and identifying RMWE is essential for achieving accurate POS tagging results. The proposed approach achieved the following evaluation metrics: a recall of 64.08%, precision of 86.84%, and an F-score of 73.74%.

Barman et al. [154] presented the application of conditional random field (CRF) and transformation-based learning (TBL) for POS tagging in Assamese phrases. They demonstrated that combining probabilistic and rule-based approaches can enhance the accuracy of the tagging process. Ojha et al. [155] discussed the training and evaluation of CRF and SVM algorithms for Indo-Aryan languages such as Hindi, Odia, and Bhojpuri. The study achieved high precision without relying on external tools but encountered challenges with proper nouns and verbs, which exhibited higher error rates. To address this issue and improve the performance of the tagger, the authors suggested incorporating Named Entity Recognition (NER) and a morph analyzer.

Ghosh et al. [156] proposed a method for POS tagging to address code-mixed text in Bengali, Hindi, Tamil, and English. The system utilizes Conditional Random Field (CRF) to capture patterns and assign accurate part-of-speech tags. Various pre-processing and post-processing modules are employed to improve system performance, yielding satisfactory results with the highest accuracy of 75.22% observed in Bengali-English mixed data.

Suraksha et al. [157] presented an approach that utilized Conditional Random Fields (CRF) for POS tagging and Chunking specifically for the Kannada language. The researchers collected a Kannada corpus of 3000 sentences from newspapers, using 2500 sentences for training and 500 sentences for testing. They reported achieving an accuracy of 96.86% for both tagging and chunking tasks.

Ajees et al. [50] presented a POS tagger based on Conditional Random Fields (CRF) specifically designed for the morphologically rich language Malayalam. In their study, they developed a CRF-based POS tagger for Malayalam and created a publicly accessible dataset with tagged annotations. The performance of their

suggested system was compared to existing approaches, and it demonstrated superior results, achieving an accuracy of 91.2%. Khan et al. [49] introduced a distinctive approach for POS tagging in Urdu, utilizing linear-chain Conditional Random Fields (CRF). This study represents the first application of CRF to Urdu POS tagging. The proposed model incorporates a balanced and robust feature set that is both language-dependent and independent, resulting in an accuracy of 88.74%.

Nasim et al. [158] presented a POS tagger for Urdu by combining two advanced models, Conditional Random Field (CRF) and Bidirectional Long Short-Term Memory CRF (BiLSTM CRF). This study introduced the first application of the BiLSTM CRF model in the Urdu POS tagging context. The results demonstrated that the BiLSTM-CRF model achieved slightly higher accuracy compared to the CRF model. The CRF model achieved an accuracy of 95%, while the BiLSTM-CRF model achieved an accuracy of 96%.

Warjri et al. [48] employed the Conditional Random Field (CRF) approach for part-of-speech (POS) tagging in the Khasi language. They created a dedicated tag set and developed a POS tagging corpus specifically tailored for Khasi. The CRF-based approach yielded favorable results, with an accuracy of 92.12%. The performance of the proposed method was compared to various other advanced techniques, demonstrating its effectiveness and promising outcomes in comparison to the alternatives considered.

5.3 Conditional Random Fields

A Conditional Random Field (CRF) is a probabilistic model that captures the relationship between a label sequence Y and an observation sequence X. The random variables representing the elements of a label sequence (denoted as Y) are conditioned on another random variable X, representing observation sequences. This conditioning implies that the label sequence Y is influenced by the observa-

tion sequence X.

A CRF can be seen as an undirected graph $G = (V, E)$, where each node $v \in V$ represents a random variable representing an element Y_v of the label sequence Y [46]. The graph G is usually defined such that it captures the dependencies between neighboring labels in the sequence. For a (Y, X) pair to be considered a conditional random field, each random variable Y_v must satisfy the Markov property with respect to the graph G. This property ensures that the label variables are conditionally independent of each other, given their neighboring variables in the graph.

The structure of graph G can be arbitrary as long as it accurately represents the conditional independencies present in the label sequences being modeled. However, when dealing with sequences, the most commonly employed type of CRF is the Linear-Chain (LC) Conditional Random Field [159][160]. In this structure, the nodes representing the elements of Y are arranged sequentially in a chain-like order. The graphical representation of a Linear Chain CRF is illustrated in Figure 5.1, where $Y_1 Y_2, ..., Y_n$ represents the output variables and $X_1 X_2, ..., X_n$ denote the observed variables, each of which can be a vector.

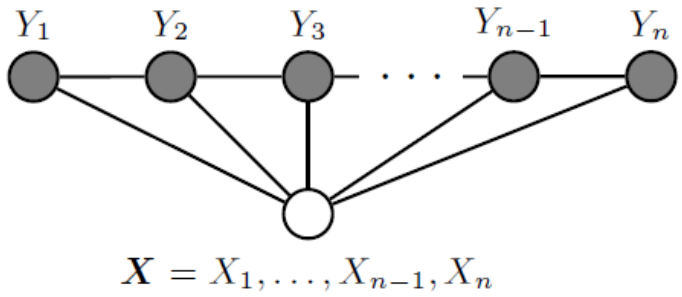

Figure 5.1: Graphical model of a Linear Chain CRF

A Conditional Random Field (CRF) is a model that directly represents the conditional distribution p(y|x), where y is a label sequence, and x is an observation sequence. The concept of CRFs for sequence labeling and data segmentation was first introduced by Lafferty et al. [46]. The probability of a particular label sequence y given the observation sequence x, is a normalized product of potential

functions. These potential functions consist of two components: transition feature functions and state feature functions, as given below:

$$exp(\Sigma_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \Sigma_k \mu_k s_k(y_i, x, i)) \tag{5.1}$$

In equation 5.1, the first part, $t_j(y_{i-1}, y_i, x, i)$, represents a set of transition feature functions, which consider the output variables at positions i and $i-1$ based on the entire observation sequence. The second part of equation 5.1, $s_k(y_i, x, i)$, is a state feature function, which takes the label at position i and the observation sequence x as inputs. The parameters of the CRF, denoted as λ_j and μ_k, are determined through training using sample data.

In the context of feature functions, a set of features denoted as $g(x, i)$ is utilized to describe these functions. These features can take the form of any real-valued positive function that captures specific characteristics of the training data, such as the empirical distribution. It is essential for the selected features to be representative of the distribution within the CRF model being considered.

So, the feature function $F_j(y, x)$ can be formulated as the summation of all individual feature functions $f_j(y_{i-1}, y_i, x, i)$. This relationship can be written as:

$$F_j(y, x) = \Sigma_{i=1}^{n} f_j(y_{i-1}, y_i, x, i) \tag{5.2}$$

Here, each $f_j(y_{i-1}, y_i, x, i)$ represents either a state function $s(y_i, x, i)$ or a transition function $t(y_{i-1}, y_i, x, i)$. This enables us to express the probability of a label sequence y given an observation sequence x as:

$$p(y|x, \lambda) = \frac{1}{Z(x))} exp(\Sigma_j \lambda_j F_j(y, x)) \tag{5.3}$$

In Equation 5.3, $Z(x)$ represents the normalization factor, which ensures that the probabilities sum to 1 over all possible label sequences for a given observation sequence x. x denotes the observed input sequence, and y signifies the output

label sequence. CRFs offer the advantage of utilizing a diverse range of features associated with each sequence element.

In the prediction phase, the inference is used to determine the most likely sequence y* for a new input x*. Dynamic-programming algorithms, such as the forward-backward algorithm and the Viterbi algorithm, can be employed to compute marginal distributions and find the most probable assignment.

5.3.1 Feature function

Feature functions are fundamental elements of a Conditional Random Field (CRF) that play a pivotal role in its structure and functionality.They are the essential components of a Conditional Random Field (CRF) as they capture the relationships between input and output variables based on patterns identified in the training data. The feature function, represented as $f(y, x, i)$, accepts the label sequence y, the observation sequence x, and the position i within the sequence as its inputs. It captures the dependencies and patterns between input and output variables at a specific position.

The specific definition and formulation of the feature function can vary depending on the task and the particular CRF model being used. Feature functions can be designed based on various factors and characteristics of the data, and they can take different forms, such as indicator functions, real-valued functions, or even more complex structures.

Let us consider the given example of a feature function defined as:

$$f_1(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & \text{if } z_n = \text{PPN and } x_n = \text{Liana} \\ 0 & \text{otherwise} \end{cases}$$

In this example, the feature function takes a pair of output states z_{n-1} and z_n, the complete input sequence $x_{1:N}$, and the position within the input sequence (n) as its inputs. The activation of this feature function depends on its weight

λ_1. If $\lambda_1 > 0$ when f_1 is active (i.e., when the word *'Liana'* is assigned a tag of PPN in the training sentence), it increases the probability of the tag sequence $z_{1:N}$. In other words, the CRF model is inclined to prefer the PPN tag for the word *"Liana"* in the test sentence. On the other hand, if $\lambda 1 < 0$, the CRF model does not favor assigning the PPN tag to *"Liana"*.

Consider another feature function based on tag transition:

$$f_1(z_{n-1}, z_n, x_{1:N}, n) = \begin{cases} 1 & \text{if } z_{n-1} = \text{PPN and } z_n = AT \\ 0 & \text{otherwise} \end{cases}$$

This feature function is activated only when there is a transition from the tag PPN (Proper Noun) to AT (Article). It is the weight λ_2 that actually specifies the log transition probability from PPN to AT.

It's important to note that feature functions in CRFs can take any real value and can be designed based on various linguistic or contextual factors. In practice, thousands of feature functions may be tailored to a particular CRF model and training set. Rather than specifying each function individually, a template is typically used to define a list of features, as shown in Table 5.2. The CRF model then generates the corresponding feature functions based on the specified features, allowing for flexibility and scalability in modeling different dependencies and patterns.

5.3.2 Inference in CRF

Conditional Random Fields (CRFs) capture the relationships between adjacent part-of-speech (POS) tags, considering the surrounding information within a sentence. Essentially, the POS tag assigned to a specific word relies not just on the word alone but also on the nearby words and their corresponding POS tags. CRFs utilize feature functions to capture these interdependencies.

During inference, CRFs calculate the probability distribution over all possible

tag sequences for a given input sentence. The goal is to find the most probable tag sequence, also known as the maximum a posteriori (MAP) sequence. The Viterbi algorithm is commonly used to efficiently find this optimal sequence.

The Viterbi algorithm works by maintaining a dynamic programming table that keeps track of the maximum probability scores for each possible tag at each position in the sentence. It proceeds iteratively, populating the table by taking into account both the transition probabilities between adjacent tags and the emission probabilities associated with the observed words, considering the corresponding tags.

Once the table is filled, the Viterbi algorithm traces back the most probable sequence by following the highest probability path. This path corresponds to the sequence of POS tags that maximizes the joint probability of the observed sentence and the tag sequence.

5.4 Experimental Set Up

The proposed system underwent several stages during the experiment, which encompassed data collection, pre-processing, tokenization, creating a tagged corpus, and configuring the software environment.

5.4.1 Data Collection and Data Preparation

The procedure for creating a resource has been detailed in the previous chapter. We take great care during the selection process to ensure the sentences in the collected texts adhere as closely as possible to grammar rules. We choose raw texts from various domains such as sports, politics, news, music, health, religion, and more. This allows us to cover a wide range of word usage across different subject areas.

We encounter several challenges during the corpus creation process. The fore-

most among these is tackling the inconsistencies in writing styles among various contributors. Many of these inconsistencies arise due to a lack of understanding of basic grammar rules among the writers. For example, take the sentence *"Ani chu a mawi ani."* The first occurrence of *'ani'* is accurately written, whereas the final instance of *'ani'* is not in accordance with the correct writing conventions. Unfortunately, there are many such instances of inaccuracy, primarily attributed to a fundamental lack of understanding regarding the writing rules of the Mizo language.

Another factor contributing to the spelling variations of Mizo words is the lack of a universally recognized standardized writing style. This poses a significant challenge during corpus development since the raw data is sourced from diverse outlets with distinct spelling conventions. For instance, writers may spell words like *'Bawng sa* (beef)' differently, with variations such as *'Bawngsa'*.

Furthermore, even within the community of linguistic experts, there are occasions where differing opinions arise regarding the proper way to write words in the Mizo language. For example, while some experts may advocate for writing *'lir nghing'* separately, others contend that it should be joined as in *'lirnghing'*. As a result, it becomes crucial to address and reconcile these discrepancies in order to create a reliable and cohesive corpus that accurately represents the Mizo language.

Another issue pertains to the generally accepted symbols for indicating different tones in the language. Currently, the symbol ‘ ̂ ’(caret or circumflex symbol) stands as the sole widely recognized symbol for denoting tones.

To ensure the accuracy and standardization of the corpus, we refer to various resources for making necessary corrections and normalizations. This involves consulting linguistic experts in the field, as well as referring to reputable grammar books [161][162][163][164]. These invaluable resources offer well-established rules and guidelines for correcting and normalizing the corpus.

5.4.2 Tagset

Chapter 4 of this research study introduced a comprehensive tagset specifically designed for Mizo language part-of-speech (POS) tagging. This tagset initially comprised 45 distinct tags as shown in Table 4.1, aiming to represent the grammatical and syntactical features of the language accurately. However, further advancements have been made in the understanding and analysis of the Mizo language, leading to the incorporation of three additional tags into the existing tagset. These new tags, namely '(, '),' and '-,' have been identified as crucial elements in capturing the intricate nuances and structures of the language. By incorporating these newly introduced tags, the tagset has become more comprehensive and robust, enabling a more precise and nuanced analysis of the grammatical features of the Mizo language. Consequently, the Mizo language POS tagset has expanded to encompass a total of 48 tags, including the three newly incorporated ones.

In fact, the hyphen (-) symbol serves a special purpose in the Mizo language, particularly when indicating hyphenated words. This is a frequent occurrence in Mizo, and during the tokenization process, the hyphenated words are treated as a single entity, with the hyphen (-) directly attached to the word. For instance, expressions like *'London-ah,' 'CYMA-in,'* and *'a lu-ah a vua'* would be tokenized as such, keeping the hyphen and the connected words intact as a single token. This approach ensures that the hyphenated words in Mizo are accurately preserved and analyzed in their entirety.

In certain cases, the hyphen (-) symbol is written separately within a sentence in Mizo. For example, in the phrase "Ka zinna lamtluang" – Rina, where it signifies that the book *'Ka zinna lamtluang'* is authored by Rina, or when listing items like *'Hengte hi kan nei - Pen, pencil, etc.'* In such instances, during the tokenization process, the hyphen is treated as a separate token, distinct from the neighboring words. This approach ensures accurate preservation of the intended meaning and structure of the sentence, allowing for precise analysis and

interpretation.

A comprehensive collection of 53,966 words has been amassed for evaluating the performance of the developed tagger. Taking into account the three supplementary tags mentioned earlier, we conducted a meticulous manual tagging process to craft a corpus with precise part-of-speech labels. This additional tag list expands the range of tags available, enabling a more comprehensive annotation of the Mizo language. This annotated corpus plays a crucial role in the advancement of research and analysis within the field of Mizo linguistics. Table 5.1 offers a summary of the tagged corpus, presenting relevant statistical information.

Table 5.1: Summary of the POS-tagged corpus

Particulars	**Count**
Total number of tokens	53,966
Total number of sentence	1656
The average number of words per sentence	32.1
Number of unique tags	48
Number of unique vocabulary	93,27
Most frequent word	A
Most frequent tag	VB

5.4.3 Specification of Features

The model has to be fed attributes for CRF feature functions, which are essentially a definition of the context of a particular word in a phrase. Table 5.2 lists the features that were chosen for the experiment. These characteristics from the training set are used by the CRF model to create feature functions.

Table 5.2: Notations for performance analysis

Sno	Features	Selected context
1	word	Current word under consideration
2	postag-1	POS tag of previous word
3	postag+1	POS tag of next word
4	is_first	First word in a sentence
5	is_last	Last word in a sentence
6	is_capitalized	First character is capitalized
7	is_all_caps	All characters are capitalized
8	is_all_lower	All characters are in lowercase
9	prefix-1	First character of a word
10	prefix-2	First two characters of a word
11	prefix-3	First three characters of a word
12	suffix-1	Last character of a word
13	suffix-2	Last two characters of a word
14	suffix-3	Last characters of a word
15	prev_word	Previous word
16	next_word	Next word
17	has_hyphen	Whether a word contains a hyphen
18	is_numeric	Whether a word consists of numbers only
19	capitals_inside	Capital letter other than first character

5.5 Experimental Results and Analysis

This section highlights the summary of results obtained from the POS tagging experiment on a custom-built corpus of 53,966 words.

5.5.1 Tagset Distribution in the Corpus

Figure 5.2 depicts a line graph illustrating the frequency distribution of the top five tags in the corpus.

As depicted in the graph, the prevailing tag evident in the corpus is the base form of a verb (VB), with a substantial count of 8,016 instances. Following this, in descending order of frequency, are the common noun (CMN), personal pronoun (PSP), proper noun (PPN), and adverb base form (RB).

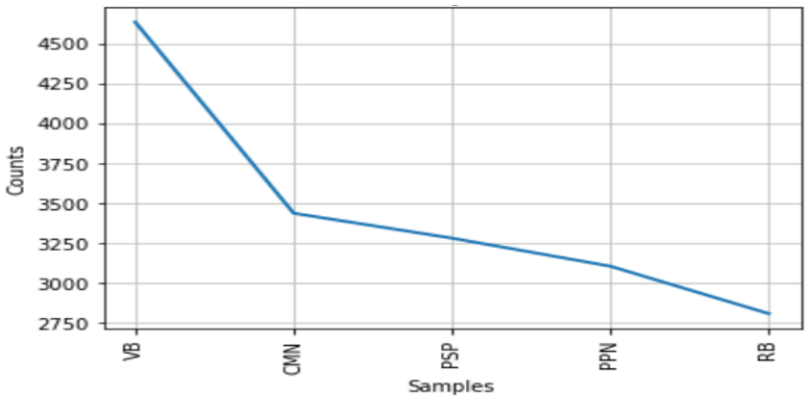

Figure 5.2: Frequency distribution of five most frequently used tags in the corpus

5.5.2 Transitions and Weights Learned by the Model

The Conditional Random Field (CRF) model learns the transitional relationship between tags in the training corpus and assigns weights accordingly. These weights reflect the strength of the relationship between consecutive output sequences. Unlike other transitions, a CRF specifically considers the transitional relationship of each consecutive output/target sequence in the training corpus. Tabulated information in Figure 5.3, the transitional weights learned by the CRF model are highlighted, focusing on the top 15 most frequent transitions in the training corpus. Tags with higher transition probabilities are assigned greater weights in the model.

From \ To	DRB	DJJ	ET	VB	IP	JJ	NG	SJJ	PT	CJJ	PT	;	PPN	AT	MP
DRB	3.743	0.0	0.0	-0.805	0.0	-0.421	0.175	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DJJ	0.0	3.219	0.0	0.0	0.0	0.115	0.0	0.0	0.0	0.0	0.0	0.0	-0.583	0.0	0.832
ET	0.0	0.0	2.682	0.0	0.0	-0.219	0.0	0.0	0.0	0.0	0.0	0.0	-0.811	0.0	0.0
VB	0.986	-0.555	-0.031	-0.514	-0.729	-0.183	2.291	-0.669	0.718	-0.798	0.718	0.0	0.168	0.012	-0.269
IP	0.0	0.0	0.0	-0.146	0.0	0.0	2.168	0.0	0.606	0.0	0.606	0.0	-0.007	0.0	0.0
JJ	1.022	-0.302	0.0	-0.214	0.233	0.098	1.283	1.829	-0.11	2.207	-0.11	0.0	0.341	0.402	0.0
NG	0.0	0.0	0.379	-1.542	0.0	0.199	0.0	0.0	0.683	0.003	0.683	0.0	-1.827	0.194	-0.033
SJJ	0.0	0.0	0.0	0.52	0.0	-0.363	0.0	0.0	-0.553	0.0	-0.553	0.0	0.0	0.4	0.0
PT	0.0	0.0	0.0	-0.404	0.0	-0.27	-0.083	0.0	1.223	0.0	1.223	2.015	-0.718	0.0	0.0
CJJ	0.0	0.0	0.0	-0.144	0.0	-0.082	0.0	0.0	0.474	0.0	0.474	0.0	0.0	0.0	0.0
PT	0.0	0.0	0.0	-0.404	0.0	-0.27	-0.083	0.0	1.223	0.0	1.223	2.015	-0.718	0.0	0.0
;	0.0	0.0	0.0	0.0	0.0	-0.399	0.0	0.0	0.0	0.0	0.0	0.0	0.276	0.0	0.595
PPN	0.0	-0.893	-0.211	0.404	0.0	0.429	-0.986	0.0	-0.381	0.0	-0.381	0.0	1.421	1.541	0.0
AT	0.0	0.0	0.329	-0.07	0.0	0.061	-2.728	0.0	-0.503	0.0	-0.503	1.379	0.476	-0.478	-0.24
MP	0.0	0.594	0.0	0.0	0.0	-0.418	0.0	0.0	0.388	0.0	0.388	0.0	0.026	-1.882	1.903

Figure 5.3: Transition weights between tags of top 15 likely transitions. (Indicated by dark green cells)

Based on the information presented in Figure 5.3, the CRF model learned

that if a word is tagged as a Double Adverb (DRB), there is a high likelihood that it will be followed by another Double Adverb (DRB).

Tabular representation in Figure 5.4 presents the list of the top 20 unlikely transitions found in the training corpus. According to the information presented

,	->	AT	-1.109799
RBT	->	AT	-1.115141
PSP	->	,	-1.177309
DRB	->	CMN	-1.185390
MJJ	->	SPRB	-1.190245
,	->	CD	-1.194791
JJ	->	CRB	-1.220804
CMN	->	NG	-1.235070
CMN	->	SRB	-1.240634
NG	->	VB	-1.282320
CC	->	.	-1.423443
,	->	PPT	-1.510990
RB	->	MP	-1.628811
CC	->	SPRB	-1.677160
NG	->	PPN	-1.712406
CC	->	PT	-1.787845
MP	->	AT	-1.851090
PSP	->	;	-2.305428
MP	->	MJJ	-2.403327
AT	->	NG	-2.435581

Figure 5.4: Top 20 unlikely transitions

in Figure 5.4, it is evident that within the training set, transitions from Article (AT) to a negation (NG) are highly improbable. These unlikely transitions are represented by negative weights, indicating that they are considered impossible based on the training corpus.

5.5.3 Feature Selected by the Model

The tabulated information provided in the following figures showcases the top features selected by the CRF model for certain feature functions, along with their corresponding assigned weights. The top features chosen for tags such as Abstract Noun (ABN), Article (AT), Co-ordinating Conjunction (CC), and Cardinal Number (CD) are shown in Figure 5.5. The top features selected for tags

y=ABN top features		y=AT top features		y=CC top features		y=CD top features	
Weight?	Feature	Weight?	Feature	Weight?	Feature	Weight?	Feature
+5.609	suffix-2:na	+2.414	prefix-3:chu	+3.537	word:leh	+4.116	is_numeric
+3.293	suffix-3:kna	+2.290	prev_word:ko	+3.232	prefix-2:ti	+3.225	is_all_caps
+3.150	suffix-3:nna	+2.273	word:chuan	+3.040	suffix-3:uan	+2.280	prev_word:nuai
+2.740	suffix-3:hna	+2.211	suffix-3:uan	+3.026	prev_word:lamte	+2.110	suffix-2:li
+2.540	next_word:chhung	+2.148	word:chu	+3.006	word:chuan	+2.097	suffix-3:nih
+2.526	prev_word:MPCPSR	+1.974	suffix-3:chu	+2.774	prefix-3:ava	+2.054	suffix-3:hum
+1.968	prefix-3:mam	+1.904	suffix-2:hu	+2.774	prefix-2:av	+1.959	next_word:%
+1.966	word:mamawhte	+1.872	prev_word:lam	+2.741	suffix-3:hse	+1.897	word:thum
+1.960	suffix-3:mna	+1.809	postag-1:PPN	+2.542	prev_word:tawng	+1.847	word:khat
... 319 more positive ...		+1.643	prev_word:lai	+2.536	suffix-3:tih	... 246 more positive ...	
... 86 more negative ...		... 114 more positive ...		... 382 more positive ...		... 53 more negative ...	
-3.193	prefix-2:hn	... 31 more negative ...		... 64 more negative ...		-2.073	postag-1:ET

Figure 5.5: Top features for ABN, AT, CC, and CD

such as Demonstrative Adverb(MRB), Negation(NG), Possessive Pronoun(POP), and Personal Pronoun(PPN) tags Figure 5.6.

y=MRB top features		y=NG top features		y=NVB top features		y=POP top features		y=PPN top features	
Weight?	Feature	Weight?	Feature	Weight?	Feature	Weight?	Feature	Weight?	Feature
+3.037	next_word:tam	+5.489	prefix-2:lo	+2.464	next_word:tuk	+3.213	next_word:pute	+3.979	is_capitalized
+3.008	prev_word:vei	+2.609	prefix-3:loh	+1.191	word:sorkar	+3.083	next_word:nu	+3.509	capitals_inside
+2.704	next_word:pawi	+2.114	prefix-1:l	+1.182	suffix-3:kar	+3.056	next_word:chhungkua	+3.385	postag-1:PPN
+1.989	prev_word:hrintir	+1.521	postag-1:IP	+1.126	prefix-3:sor	+2.848	next_word:pa-in	+3.244	next_word:ti
+1.638	next_word:chhuk	+1.387	prev_word:tha	+1.125	prefix-2:so	+2.469	next_word:pa	+3.129	next_word:gram
+1.520	next_word:kir	+1.272	suffix-3:mah	+1.101	postag+1:RB	+2.209	next_word:khaw	+2.979	next_word:'
+1.504	word:han	+1.254	word:mah	+1.050	suffix-2:ar	+2.083	next_word:nunna	+2.576	next_word:motor
+1.502	prefix-3:han	+1.233	postag-1:VB	+1.041	prev_word:pawh	+1.766	next_word:state	+2.540	postag+1:PPN
+1.458	suffix-1:o	... 70 more positive ...		+1.036	prev_word:Hawla	+1.666	next_word:kawngpui	+2.501	prefix-3:Lal
+1.261	word:kha	... 12 more negative ...		+0.992	prefix-1:s	+1.648	prev_word:atanga	... 1495 more positive ...	
... 112 more positive ...		-1.343	prev_word:thlen	... 13 more positive ...		... 83 more positive ...		... 390 more negative ...	
... 4 more negative ...		-1.826	postag+1:VB	... 2 more negative ...		... 4 more negative ...		-3.052	suffix-2:-a

Figure 5.6: Top features for SPRB, SRB, SYM, UH, and VB

The provided tabulated data in the figures demonstrate that the features chosen by the CRF model, along with their assigned weights offer a reliable means of categorizing words into probable tags. Furthermore, it indicates the significant influence of word context in determining the appropriate tag for a given word.

For instance, consider a feature selected for Abstract Noun(ABN). The feature 'suffix-2:na' (The last two characters of a word is 'na') is given a high weight value of 5.609. This is an accurate selection because in the Mizo language, it is observed that a significant number of words ending with *'na'* are indeed Abstract Nouns. E.g. *Huaisenna, lawmna, hlimna*

Likewise, in the case of a Cardinal Number (CD), the feature 'is_numeric' is assigned a significant weight value. This is because any value that can be identified as numeric is highly probable to be categorized as a Cardinal Number.

5.5.4 Quality Metrics Used

In this section, we will discuss the metrics that were employed to measure and evaluate the results of our experiment. The metrics used include accuracy, precision, recall, and f1-score, each of which provides valuable insights into different aspects of the performance of our approach. The expressions for these metrics are as follows:

Accuracy: Accuracy is a fundamental metric that measures the overall correctness of the tagger's predictions. It represents the proportion of correctly classified instances out of the total number of instances. The accuracy score provides a general assessment of the tagger's ability to assign the correct tags to the given data, regardless of the specific class or category. While accuracy provides a straightforward measure of overall correctness, it may not be suitable when dealing with imbalanced datasets, where the number of instances for different classes varies significantly. Accuracy can be calculated as

$$Accuracy = \frac{No\ of\ correct\ tags}{No\ of\ words} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where

TP = No. of true positives,

FP = No. of false positives,

FN = No. of false negatives,

TN = No. of true negatives

Precision: Precision is a metric that focuses on the tagger's ability to accurately predict positive instances. It measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive (true positives + false positives). Precision provides insights into the tagger's ability to accurately identify and assign the correct tags when it predicts a positive instance. A higher precision score indicates a lower rate of false positive predictions,

implying a higher level of confidence in the correctness of positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall, also known as sensitivity, gauges the tagger's ability to capture all actual positive instances. It measures the proportion of correctly predicted positive instances (true positives) out of all actual positive instances (true positives + false negatives). Recall evaluates the tagger's ability to capture all the relevant tags in the data, ensuring that it doesn't miss any positive instances. A higher recall score indicates a lower rate of false negative predictions, suggesting that the tagger can effectively identify the majority of the positive instances in the data.

$$Recall = \frac{TP}{TP + FN}$$

F1-score: The F1-score is a composite metric that balances precision and recall. It provides a single measure that combines these two metrics to evaluate the overall performance of the tagger. The F1-score is calculated as the harmonic mean of precision and recall, providing an assessment that considers both the ability to identify positive instances accurately and the capability to capture all actual positive instances. It is particularly useful when dealing with imbalanced datasets or when both precision and recall are important.

$$F1 - Score = 2X\frac{PrecisionXRecall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

By utilizing these metrics, we can gain a comprehensive understanding of the performance of our tagger. While the F1-Score offers a holistic evaluation by considering both precision and recall, accuracy, precision, and recall individually provide specific insights into different aspects of the tagger's performance, enabling us to analyze and improve its strengths and weaknesses.

5.5.5 Performance Reports

In order to evaluate the overall effectiveness of the CRF tagger, the corpus is divided into two sets: training and testing. Different split ratios, such as 70:30, 75:25, 80:20, 85:15, and 90:10, are used for the train and test sets, respectively. The performance of each metric is calculated using the specified formula, and the results are recorded in Table 5.3. This table provides a comprehensive overview of the performance of the CRF tagger for different split ratios. In addition to the overall performance evaluation presented in 5.3, a more detailed analysis of the CRF tagger's performance is provided in Table 5.4. This table specifically focuses on the tag-wise scores of precision, recall, and F1-score, utilizing a 90:10 split ratio for the train and test sets.

Table 5.3: Performance report

Train set : Test set	**Accuracy**	**Precision**	**Recall**	**F1-score**
70:30	84.17%	79.66%	77.14%	78.38%
75:25	84.75%	80.47%	77.68%	79.05%
80:20	85.24%	81.31%	78.11%	79.68%
85:15	85.76%	81.85%	7856%	80.17%
90:10	86.84%	82.10%	79.02%	80.53%
Average Score	85.35%	81.08%	79.02%	79.56%

The outcomes generated by the suggested POS tagger demonstrate consistent and encouraging performance across a range of assessment criteria. The mean values for accuracy, precision, recall, and F1-score are documented at 85.53%, 81.08%, 79.02%, and 79.56%, respectively. These results suggest that the POS tagger exhibits a noteworthy degree of accuracy and efficiency in categorizing and labeling parts of speech within the provided dataset.

The average F1 score, which combines precision and recall, stands at 79.56%. This score indicates a balance between the tagger's ability to correctly identify relevant instances of each part of speech (recall) and its capability to precisely assign the correct tags (precision). The high F1 score implies that the tagger per-

Table 5.4: Precision, recall, and F1-score for each tag

Tags	Precision	Recall	F1-score	Tags	Precision	Recall	F1-score
ABN	0.88	0.94	0.91	ET	0.86	0.74	0.80
PSP	0.97	0.84	0.98	QM	1.00	1.00	1.00
VB	0.92	0.93	0.92	DRB	0.78	0.92	0.84
RB	0.91	0.87	0.89	NG	0.99	1.00	0.99
CC	0.88	0.78	0.92	POP	0.92	0.82	0.87
JJ	0.94	0.86	0.65	SJJ	0.88	0.88	0.88
SYM	1.00	1.00	1.00	(	1.00	1.00	1.00
FW	0.90	0.82	0.86	)	1.00	1.00	1.00
.	1.00	1.00	1.00	MP	0.82	0.70	0.76
PPN	0.85	0.89	0.87	CJJ	0.60	1.00	0.75
CMN	0.84	0.86	0.85	MRB	0.81	0.84	0.82
MJJ	0.94	0.84	0.89	-	1.00	1.00	1.00
RBP	0.81	0.83	0.82	:	1.00	1.00	1.00
PPT	0.91	0.83	0.87	IP	0.77	0.55	0.64
DJJ	1.00	1.00	1.00	NVB	0.00	0.00	0.00
,	1.00	1.00	1.00	CRB	1.00	0.5	0.67
RBM	0.94	0.82	0.88	SRB	0.86	0.86	0.86
AT	0.98	0.74	0.84	SF	0.00	0.00	0.00
CD	0.97	0.75	0.85	UH	0.00	0.00	0.00
SPRB	0.67	0.84	0.75	?	1.00	1.00	1.00
PT	0.98	0.76	0.86	VBN	0.88	0.76	0.82
;	1.00	1.00	1.00	IJJ	0.62	0.84	0.71
RLP	0.00	0.00	0.00	IRB	0.68	0.72	0.70
RBT	0.79	0.82	0.80	DVB	0.86	0.78	0.82

forms well in both these aspects, resulting in accurate and reliable POS tagging.

The POS tagger's performance is notable, particularly in terms of precision (81.08%) and recall (79.02%). This indicates that the tagger excels in accurately identifying and classifying parts of speech. The balance between precision and recall suggests a well-tuned model that strikes a favorable trade-off between minimizing false positives and false negatives, making it a valuable tool for accurate linguistic analysis.

The remarkable outcomes showcase how well the suggested POS tagger performs in precisely labeling words, facilitating precise syntactic and semantic analysis of the provided text. Such high scores across multiple evaluation metrics are indicative of the tagger's robust performance and its potential to enhance vari-

ous natural language processing tasks that rely on accurate POS tagging. Based on the current results, the tagger exhibits a commendable level of accuracy and precision, making it a valuable tool for a range of linguistic and computational applications in Mizo.

5.6 Analysis on External Sentences

We conducted a performance comparison with the Hidden Markov model (HMM) tagger, which was discussed in the previous chapter. This comparison involved the use of three external sentences specifically designed to assess the taggers' abilities in various situations. These test sentences deliberately include several instances of ambiguity caused by factors such as words with identical spellings but distinct tones and meanings, words with identical spellings and tones but different meanings, and different parts of speech.

Below, the test sentences are provided, along with the actual tags, as well as the tags generated by both the HMM-based model and the CRF model. Each word is followed by the assigned tag, separated by a slash (/).

Test sentence 1: *Aizawlah ka kal dawn*
Actual tag : Aizawlah/RBP ka/PSP kal/VB dawn/RB
HMM : Aizawlah/NN ka/PSP kal/VB dawn/RB
CRF : Aizawlah/NN ka/PSP kal/VB dawn/RB

Test sentence 2: *Kan inah min lo kan rawh*
Actual Tag: Kan/POP inah/RBP min/PSP lo/MRB kan/VB rawh/PT
HMM : **Kan/PSP** inah/RBP min/PSP **lo/NG kan/PSP** rawh/PT.
CRF : Kan/POP inah/RBP min/PSP lo/NG kan/VB rawh/PT.

It is observed that both the HMM and CRF models demonstrate accurate tagging of words in the first sentence. In the second sentence, we encounter an

ambiguous word, *'kan,'* which appears twice. In the first occurrence, it signifies 'our,' and in the second occurrence, it signifies 'visit.' While the CRF model correctly assigns tags to all words in this sentence, the HMM model encounters difficulties and incorrectly tags the second instance of *'Kan,'* failing to distinguish its intended meaning as 'visit.'

Test sentence 3: *Chu ri chu, lei piah zawk ka zawh zawha thil ka lei a piangin a ri zawi thin.*

Actual: Chu/RLP ri/CMN chu/RLP, lei/CMN piah/JJ zawk/CJJ ka/PSP zawh/VB zawha/RB thil/CMN ka/PSP lei/VB a/PSP piangin/PPT a/PSP ri/VB zawi/RB thin/RB

HMM : Chu/AT ri/CMN **chu/AT, lei/VB piah/RB zawk/ABN** ka/PSP zawh/VB **zawha/RBT** thil/CMN ka/PSP **lei/CMN** a/PSP piangin/PPT a/PSP ri/VB zawi/RB thin/RB

CRF : Chu/AT ri/CMN **chu/AT, lei/VB piah/RB** zawk/CJJ ka/PSP zawh/VB zawha/RBT thil/CMN ka/PSP lei/VB a/PSP piangin/PPT a/PSP ri/VB zawi/RB thin/RB

In the third sentence, an examination of part-of-speech tagging reveals that the HMM assigned incorrect tags to the words ***'chu,' 'lei,' 'zawk,' 'zawha,'*** and ***'piah.'*** Notably, when employing the CRF model, an improvement in tagging accuracy is evident. Specifically, the words 'lei,' 'zawha,' and 'piah' have been correctly assigned their respective tags. Nevertheless, issues persist with words such as ***'lo,' 'chu,'*** and the initial occurrence of ***'lei,'*** which continue to be mislabeled.

Overall, one potential reason for these tagging errors could be attributed to an imbalance in the dataset. For example, the word *'chu'* has been labeled as 'AT' 367 times, while it has only been labeled as 'RLP' 7 times. This emphasizes the significance of having a comprehensive dataset that encompasses a wide range of possibilities and variations in language usage.

Additionally, the nature of the Mizo language, being tonal, presents another

challenge for accurate tagging. Take, for example, the word *'lei.'* Depending on its tone, it can carry different meanings. The high tone *'lei'* and the low tone *'lei'* have distinct interpretations. This tonal aspect of the language can pose difficulties for the tagger in accurately assigning the appropriate tags, leading to incorrect results.

5.7 Conclusion

Within this chapter, we have outlined an approach based on Conditional Random Fields (CRF) for the automated POS tagging of Mizo text, yielding an impressive average accuracy of 85.35%. Despite its simplicity, the CRF model presented herein proves to be a potent tool for automatic tagging, particularly in scenarios where labeled text resources are limited. The development of a tagged corpus, encompassing 53,966 words stands as a significant stride in advancing this low-resource language. As we move forward, our objectives encompass the expansion of the tagged corpus and the assessment of alternative language models to enhance the efficacy of our approach.

Chapter 6

HYBRID POS TAGGER

6.1 Introduction

In recent years, part-of-speech (POS) tagging has become an essential task in natural language processing (NLP) applications. POS tagging involves assigning a unique grammatical tag to each word in a sentence, indicating its syntactic and semantic role in the sentence. Several approaches have been proposed for POS tagging, including rule-based taggers, N-gram taggers, Hidden Markov Model (HMM)-based taggers, Conditional Random Field (CRF)-based taggers, and different deep learning models.

In the previous chapters, we have explored two prominent methodologies for Part-of-Speech (POS) tagging in the Mizo language: Hidden Markov Model (HMM) and Conditional Random Field (CRF). These models have demonstrated their efficacy in Mizo POS tagging, yielding promising outcomes. However, it is crucial to acknowledge their inherent limitations. One common issue is that both the taggers heavily depend on labeled training data. They may struggle when faced with domain-specific or low-resource scenarios where obtaining sufficient labeled data is challenging.

In this chapter, we present novel methods for enhancing the performance of two popular POS tagging techniques: the N-gram backoff tagger and the Hidden Markov Model (HMM)-based tagger. We address the limitations of these approaches by incorporating rule-based techniques, modifying the decoding algorithm, and integrating different taggers to create a hybrid system. The objective of our research is to improve the accuracy, robustness, and adaptability of the Mizo POS tagging systems.

The first part of this chapter focuses on enhancing the N-gram backoff tagger using rule-based techniques. The N-gram backoff tagger is known for its simplicity and efficiency, but it often struggles with handling ambiguous words and rare or unseen words. To overcome these challenges, we propose a method that integrates rule-based strategies into the N-gram model. By capturing linguistic patterns and context-specific rules, we aim to improve the accuracy of tagging decisions, particularly for complex cases. Our approach combines the statistical strength of the N-gram model with the linguistic knowledge embedded in the rule-based techniques.

Next, we introduce a novel and simple modification to the decoding algorithm (i.e. Viterbi algorithm) used in the HMM-based POS tagger, specifically targeting the scenario where the model encounters words that were not present in the training set. In conventional HMM-based taggers, the Viterbi algorithm tends to assign default tags to unseen words, which often leads to incorrect tagging outcomes. To tackle this problem, we propose a straightforward method for estimating the most probable tag for unseen words. Through rigorous experimentation, we have successfully demonstrated the effectiveness of this approach in improving overall performance.

Finally, we propose a hybrid approach for POS tagging that combines the strengths of three different taggers: the HMM-based tagger, the rule-based tagger, and the N-gram tagger. The hybrid POS tagger uses a combination of HMM-based, rule-based, and N-gram models to assign tags to words in a sentence. The proposed hybrid tagger is designed specifically for the Mizo language, and the main objective of this proposed work is to investigate the potential for improving the efficiency of the HMM-based tagger designed for the Mizo language.

In the subsequent sections, we delve into the methods employed to enhance the N-gram taggers, followed by a detailed examination of the individual components comprising the hybrid POS tagger. These components include a regular expression tagger, the HMM-based tagger, and the N-gram model. We also elu-

cidate the process of combining these models to create the hybrid POS tagger. Lastly, we present the experimental results and conduct a comparative analysis of the performance of the proposed hybrid tagger with that of the HMM and N-gram taggers.

6.2 Related Works

Several academics have already begun working on building POS taggers, employing various algorithmic approaches. In recent years, considerable effort has been expended on the POS tagging of Indian languages too. Some of the POS tagging works done for various Indian languages are highlighted below.

One of the earliest articles on POS taggers for the Hindi language was published by Ranjan et al. [100]. It was based on the lexical tags of words. Many new approaches have emerged since the works were first published. Modi et al. [165] presented a paper on a hybrid POS tagger for Hindi. The model combined a rule-based and a probability-based model. The model was trained with 9000 words and yielded an average accuracy of 88.15%. The attention-based model for POS tagging was presented by Mundotiya et al. [112]. The model was tested on the Hindi disease dataset, and an accuracy of 98.36% was reported.

A Hindi POS tagger based on the Maximum Entropy Markov Model was developed by Dalal et al. [44]. Multiple features were employed simultaneously in this model to predict the word's tag. The model was trained with 3500 words annotated with 29 different tags and the average accuracy reported by the model was 88.4%. Applying Artificial Neural Networks, Narayan et al. [166] built a POS tagging system for Hindi. The proposed model was compared to other strategies, such as Maximum Entropy-based, Rule-based, and CRF-based taggers. The experimental findings demonstrated that the proposed tagger outperformed all of them.

A statistical POS tagger for Bengali using the Maximum Entropy (ME) model

was presented by Ekbal et al. [45]. On a testing dataset of 20,000 words, the POS tagger was trained with 72,341 words and obtained an overall accuracy of 88.2%. Experiments have shown that the different word suffixes, named entity recognizer, and lexicon can help with unknown word difficulties and considerably enhance the POS tagger's accuracy significantly. Jahara et al. [167] presented the results of an empirical study of various POS tagging approaches for the Bengali language. Among all the tagging approaches, Brill coupled with CRF had the best accuracy of 91.83 %.

Sharma et al. [168] proposed a bi-gram HMM-based POS tagger for the Punjabi language. The model was put through its paces on a corpus of 26,479 tokens. Using this procedure, an accuracy of 90.11% was reported. Jobanputra et al. [169] proposed employing Long-Short-Term Memory to construct a POS tagger for the Gujarati language (LSTM). They claimed a 95.34% accuracy. Tailor et al. [170] suggested a hybrid technique for the Gujarati POS tagger, consisting of the Computational Linguistic rules and the LSTM-based POS tagging. The experimental outcome indicated that adopting language-specific rules improved the statistical tagger. The review paper by Gamit et al. [171] presented various methods of POS tagging in the Gujarati language.

A POS tagger for the Maithili language based on the CRF model was proposed by Priyadarshi et al. [172]. They built a corpus containing around 52K words, which were manually annotated with the designed tagset. They have experimented with various orthography features in the model and achieved an accuracy of 82.67%. Swamy et al. [47] made a POS tagger for the Kannada language by taking the CRF as their model. A corpus consisting of 234k words was employed for the experiment and the test results reported 91.4%, 91.6%, and 91.3% for the values of f-score, recall, and precision, respectively.

Assamese POS tagging based on HMM was presented by Daimary et al. [173]. A corpus consisting of 271,890 words was utilized in the experiment and achieved 89.21% accuracy. Another POS tagger for the Assamese language using a deep

learning method was presented by Pathak et al. [174] and obtained 86.52% accuracy. Singh et al. [153] developed Manipuri POS taggers using CRF and SVC. The models were trained with 39,449 tokens and tested with 8,672 tokens. They have obtained accuracies of 72.04% and 74.38% for the CRF and SVC models, respectively.

A hybrid POS tagger for Khasi was developed by Tham [175]. In this paper, an HMM-based tagger was integrated with the CRF model yielding an accuracy of 95.29%. Another Khasi POS tagger using Deep learning methods was presented by Warjri et al. [176]. They developed the BiLSTM approach, which was integrated with the CRF and character-based models. According to the testing data, the BiLSTM combined with the CRF model produced a maximum accuracy of 96.98%. Vaishali et al. [177] presented a Marathi POS tagging system using a rule-based technique. The experimental result demonstrated an accuracy of 97.56%.

Many more diverse POS tagging works have been highlighted in these papers [178, 179, 180]. In conclusion, various approaches, such as rule-based, machine learning, and deep learning techniques, have been extensively employed in POS tagging for Indian languages. As documented in this review, the most notable achievement in terms of accuracy stands at 98.36% [112]. Moreover, publicly available POS-tagged corpora for several Indian languages have been published [181].

Research in Mizo language POS tagging is notably limited, with Pakray et al.'s work [117] standing as the sole paper in this domain (to the best of our knowledge). Their study primarily focused on resource development, resulting in a Mizo-to-English dictionary and a dedicated part-of-speech (POS) tagger. This effort involves amassing 26,407 entries and creating a 24-item POS tag set. Importantly, this work paves the way for an automated POS tagging system for Mizo, addressing a critical deficiency in linguistic resources for this underrepresented language.

In summary, our exploration of related work reveals a diverse landscape of POS tagging methods applied to various Indian languages. However, it is worth noting that the field of language processing, particularly POS tagging, remains underrepresented in the context of the Mizo language. This underscores the significance of our current efforts to contribute to the advancement of linguistic analysis in the Mizo language through the development of effective POS tagging models.

6.3 Building a Regular Expression tagger

A regular expression tagger is a type of rule-based tagger used in part-of-speech (POS) tagging. It works by using a set of regular expressions to match patterns in the text and then assigns a POS tag to the word based on the matched pattern.

After examining and analyzing the structure of the morphologically rich Mizo language, a set of regular expressions has been designed with the assistance of language experts and the authoritative Mizo grammar book publications [182] [183]. It was carefully crafted due to the sophistication and grammatical complexity of the Mizo language. The selection and design of the regular expressions were specifically tailored to the morphological structure of the Mizo language, taking into consideration the unique characteristics of the language.

The purpose of using a regular expression tagger is to improve the accuracy of the tagging process, especially for words that are difficult to tag using our proposed statistical models. It is used to capture specific morphological or syntactic patterns in the text, such as suffixes or prefixes, and assign appropriate tags based on these patterns. For instance, if the regular expression finds a word ending with *"na"* e.g. *"hmangaihna", "lawmna", "remna",* it will tag each of them as "CMN(Common Noun)".

Because this analysis focuses solely on individual words and ignores their context within sentences, the generated results may not provide accurate insights for

all terms in every situation. To illustrate this point, take the following sentence into consideration:

"Kah rawn a hlawh"

The word *"Kah"* in the above line is a "Verb". However, since the regular expression tagger finds that the word ends with *"ah"* it will wrongly tag it as "RBP (Adverb of Place)."

In addition to the trivial tagging of elements such as symbols, conjunctions, articles, and personal pronouns, the analysis and the clues are provided below:

- Foreign Word(FW) :

  - If a word contains characters that are not available in Mizo alphabets, such as Q(q), X(x), Y(y)
  - If a word contains W(w) but not prefixed by A(a)
  - If a word contains a character H(c), but not followed by H(h)
  - If a word contains a character G(g), but not prefix by N(n)

- Proper Noun(PPN) :

  - If a word starts with a capital letter and ends with 'i' or 'a'
  - If all letters of a word are capital
  - If a word ends with '-in'

- Abstract Noun(ABN) :

  - If a word ends with 'na'

- Common Noun(CMN) :

  - if a word ends with 'te' or 'ten'

- if a word ends with 'ho'
- if a word ends with 'pui'
- if a word ends with 'tu'
- if a word ends with 'in'

-Demonstrative Pronoun(MP)

- If a word ends with 'ngte'

- Adverb of Place (RBP)

- If a word consists of characters only and ends with '-ah' or '-a'
- If a word ends with 'ah'

- Adverb of Time (RBT)

- If a word consists of digits only and ends with 'ah' '-a'

-Verb base form (VB)

- if a word starts with 'In' or 'IN'
- if a word ends with 'san'
- if a word ends with 'tir'

-Date (ET)

- if a word is a month or day of the week

-Adjective base form (JJ)

- if a word ends with 'zia'

-Default tag: VB

6.4 Improving the N-gram Backoff Tagger

The N-gram model is one of the most straightforward language models for assigning probabilities to phrases and word sequences. It is a collection of n words, with the number 'N' indicating the length or size of the words. For example, size one N-gram is referred to as a unigram or 1-gram, size two N-gram is called a bigram or 2-gram, and size three N-gram is referred to as a trigram or 3-gram, and so on.

The initial step of the tagging process involves constructing a context for each token in order to determine the appropriate tag for it. This context is made up of the token type and part-of-speech tags of the 'N' tags that came before it. The N-gram POS taggers choose tags by considering the word sequence and the tags of the n preceding words. These N-gram taggers rely on a training corpus with labeled tags to identify the most probable part-of-speech tag for each specific context. Each tagger maintains a context dictionary, which is utilized to make an educated guess for a tag based on the context. Figure6.1 depicts an example of an N-gram tagger with n=3, where the context for determining the tag t_n is tinted gray that includes t_{n-2}, t_{n-1}, and w_n.

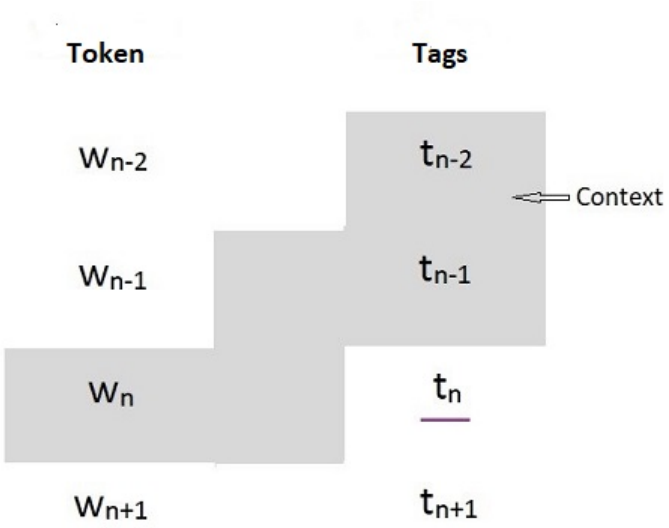

Figure 6.1: Context of the N-gram tagger

In the N-gram model, the level of context specificity also increases as 'N'

increases. However, this can lead to a higher likelihood of encountering data that lacks context information within the training corpus. In natural language processing (NLP), this problem is known as the sparse data problem, and it poses challenges in various research areas.

The sparse data problem becomes particularly evident when using trigram models, as the instances generated with trigrams alone are often insufficient for directly calculating the probability in the tagging system. The underlying principle is that when working with limited training data, it becomes challenging to observe the true distribution of word occurrences in the source language due to the scarcity of training samples.

To address this issue in our approach, we have employed the backoff tagging technique. This technique allows the system to "fallback" to lower-order N-gram models (e.g., bigrams or unigrams) when faced with insufficient data for a specific trigram context. By utilizing lower-order models as a backup, we can mitigate the impact of sparse data and improve the accuracy of our tagging system.

In the backoff tagging approach, when a tagger encounters a word for which it doesn't have any tagging information, it can delegate the task to the next backoff tagger in the hierarchy. If the subsequent tagger is unable to complete the tagging, it can pass on the task to the next tagger in line, continuing this process until there are no more backoff taggers left to check.

To handle ambiguous terms within the corpus and enhance the effectiveness of N-gram taggers, we have integrated a designed regular expression tagger into the N-gram backoff tagging system. This integration helps improve the efficiency of the N-gram taggers in dealing with ambiguous terms. Specifically, the trigram backoff tagger first attempts to tag a word using the trigram tagger by searching for the sequence of trigram probabilities within the training data. If the trigram tagger cannot find a suitable sequence, it falls back to the bigram tagger. If the bigram tagger also fails to find a probability for the word sequence in the training data, the model further falls back to a unigram tagger. If even the unigram

tagger is unable to tag the word, the word is then passed to the designed regular expression tagger for further consideration and potential tagging.

6.5 Enhancing an HMM tagger

This section discusses two methods for improving the effectiveness of an HMM-based tagger. The first method involves tweaking the decoder used, while the second method involves integrating the HMM-based tagger with both the N-gram tagger and the regular expression tagger.

6.5.1 Method I: Modifying the Decoder

One of our proposed methods for enhancing the performance of an HMM-based tagger specifically designed for the Mizo language is presented in this section. Our approach involves making modifications to the decoder component.

The standard approach of using a Hidden Markov Model (HMM) to find the most probable tag sequence is computationally expensive, as the time required to solve the problem grows exponentially. To address this issue, we employed the Viterbi method, which is a widely used decoding algorithm for HMM-based part-of-speech tagging. The Viterbi algorithm, based on dynamic programming, aims to determine the most likely sequence of hidden or unobservable states, known as the Viterbi path, that corresponds to a series of observed events. The Viterbi algorithm takes a set of observations, $W = (w_1 w_2 w_3 w_4 ... w_n)$ as an input and returns the most likely state sequence, $S = (s_1 s_2 s_3 s_4 ... s_n)$ with its probability.

The Viterbi algorithm involves the creation of two probability matrices: one for transition probabilities and one for emission probabilities. During the testing phase, this Viterbi decoding algorithm utilizes these matrices of tag transition probabilities and emission probabilities to calculate the most probable sequence of tags for each phrase in the input corpus.

In this work, the baseline version of the Viterbi algorithm is modified to handle unknown terms in the corpus, resulting in improved performance of the HMM-based tagger. As mentioned earlier, the Viterbi algorithm utilizes both emission probabilities and transition probabilities to determine the most probable tag sequences. However, if a word is not found in the training data (meaning it is an unknown word), its emission probability becomes zero, resulting in a state probability of zero. In such cases, when the algorithm encounters an unseen word during training, it exclusively relies on the transition probability to calculate the state probability, disregarding the emission probability. Thus, the algorithm operates in the following manner:

if $Word\ Not\ In\ Vocabulary$ **then**

$State\_prob = Trans\_prob$

else

$State\_prob = Trans\_prob * Emi\_prob$

end if

The algorithm mentioned above can be interpreted as follows: If a word is not present in the vocabulary, then the state probability is equal to the transition probability; otherwise, the state probability equals the product of the transition probability and the emission probability.

With this simple method, the baseline version of the HMM-based tagger for the Mizo language has been improved, allowing it to produce better results. Nonetheless, the exclusive reliance on transition probabilities for out-of-vocabulary (OOV) words in POS tagging comes with inherent limitations. It leads to a loss of valuable word-specific information and ignores contextual cues, which are essential for accurate tagging. Additionally, its effectiveness heavily relies on the quality of training data, potentially yielding suboptimal results if the data is inadequate or lacks comprehensive transition probabilities for OOV words.

6.5.2 Method II : Developing a Hybrid POS Tagger

This section describes the proposed method for enhancing the HMM-based POS tagger for the Mizo language by integrating the hidden Markov model (HMM) with the N-gram model bigram tagger from the NLTK package and the rule-based tagger. The resulting tagger will be referred to as a hybrid tagger.

The HMM-based tagger forms the backbone of our hybrid approach. It models the probability distribution over sequences of states, where each state corresponds to a particular POS tag. However, HMM-based taggers may fail to accurately tag words that are not present in the training corpus or have ambiguous contexts.

The N-gram tagger uses the POS tags of the N preceding words (where N is customizable) to determine the probability of a tag being assigned to the current word. This probability is based on the frequency of the tag in the context of the preceding N words. Based on these probabilities, the tagger assigns the most likely tag to the current word.

The rule-based tagger allows for the explicit encoding of linguistic patterns and domain-specific knowledge to guide the tagging process. This tagger relies on handcrafted rules that capture the Mizo language's specific morphological, syntactic, and semantic patterns. By incorporating these rules, the hybrid tagger can handle cases where statistical models may fall short, providing an additional layer of accuracy and robustness. Rule-based taggers can be very accurate, but they require significant time and effort to create and may not generalize well.

By combining the strengths of the HMM, the N-gram model bigram tagger, and the rule-based tagger, the proposed hybrid Mizo POS tagger presents a comprehensive solution for accurate and efficient tagging of parts of speech in the Mizo language. These three components are seamlessly integrated within a well-defined framework, ensuring a coherent and systematic approach to combining their outputs.

Overall Architecture of the Hybrid Tagger

The hybrid POS tagging model presented in this research combines the strengths of Hidden Markov Models (HMM), N-grams, and rule-based taggers to address the complexities of part-of-speech tagging. Figure 6.2 depicts the overall architecture of the proposed tagger, which can be outlined as follows:

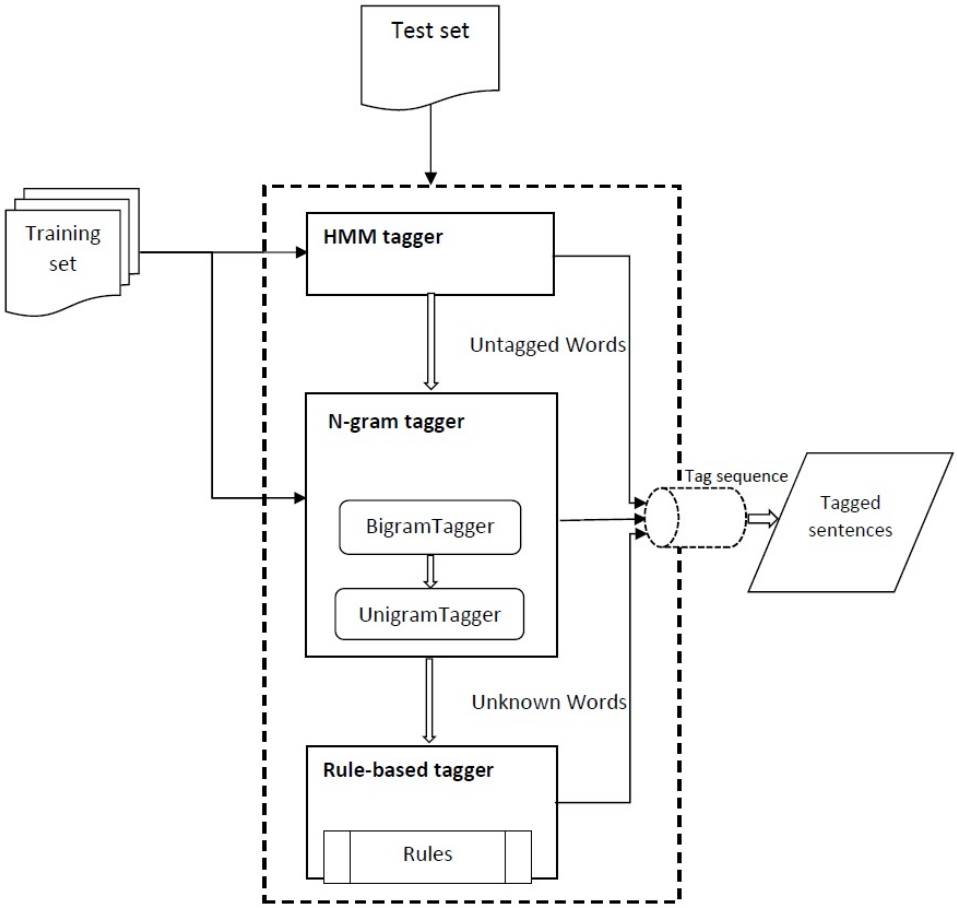

Figure 6.2: The overall architecture

Input Processing: The model begins by taking a sequence of words (a sentence or text fragment) as input. This input is tokenized into individual words or tokens, which form the basis for subsequent tagging.

Initial HMM Tagging: The first component of the model is an HMM-based tagger. It assigns initial POS tags to each word in the input sequence based on learned transition probabilities and emission probabilities. The Viterbi

algorithm is employed to perform this initial tagging. Subsequently, to determine the confidence of the HMM-based tagging, a thresholding mechanism is applied to the state probability of the tag sequence produced by HMM. If the probability falls below a predetermined threshold, it triggers further processing.

N-gram Backoff Tagging: When the HMM-based tagging is deemed uncertain (below the threshold), the model seamlessly switches to an N-gram-based tagger(bigram backoff tagger is selected for this study). The bigram tagger performs a lookup for a tuple that contains the previous tag and the current word in the context to find the appropriate tag for a given word. If the bigram fails to identify a suitable tag for the provided word, it delegates the task to the unigram tagger.

The unigram tagger employs a basic statistical technique to tag a token by selecting the most frequently encountered tag associated with that token in the annotated training text corpus. The unigram tagger will not be able to label words that are not found in its vocabulary, commonly known as out-of-vocabulary words (or unknown words). These words are then forwarded to the rule-based tagger.

Rule-based tagger: The third component introduces linguistic heuristics and rule-based patterns to the tagging process. This rule-based tagger assesses unknown words obtained from the unigram tagger and endeavors to assign them suitable tags according to predefined rules. Given that "VB (Verb)" is the most prevalent tag in the corpus utilized in this study, it has been designated as the default tag. Therefore, if the rule-based tagger is unable to identify an appropriate rule for tagging a token, the token will be assigned the default tag "VB (Verb)."

6.6 Implementation and Result Comparison

Two different experimental setups were employed to carry out the implementations and evaluate the results. The initial experiment aimed to assess the effectiveness of the suggested techniques in improving N-gram taggers. Subse-

quently, another experiment was conducted to evaluate the performance of the two techniques in enhancing an HMM-based tagger.

6.6.1 Data Used for the Experiment

The experiments were carried out using a dataset comprising 72,077 words. In order to assess and compare the performance of the taggers, a training set to test set ratio of 90:10 was adopted. Table 6.1 displays statistics of the designed Mizo corpus, consisting of 77,8037 tokens (statistics for training and testing data are provided from a single run). Once again, we noted that verbs are the most prevalent tags in this corpus, consistent with our earlier observations.

Table 6.1: Corpus Statistics

Total no. of words	77,803
No. of training sentences	2,208
No. of testing sentences	246
No. of train tagged words	69,154
No. of test tagged words	8,649
No. of unique words in train set	9,141
No. of unique tags	48

We employ a tagset of 48 tags, which was developed in the previous chapter, to annotate the collected text for subsequent processing. Table 6.2 presents the frequency of the tagset in the corpus, highlighting the most frequent tags in the dataset.

It was observed that the dataset exhibited an imbalance, as indicated in Table 6.2. To address this data imbalance, we employed the *"stratified k-fold cross-validation"* method while evaluating the performance of the taggers. This approach involved splitting the dataset into 10 equally sized "folds," ensuring that each fold contained the same ratio of instances of each tag as in the original dataset. The first nine folds were used for training during the evaluation, and the last fold was kept for testing. This process was repeated ten times, with

Table 6.2: List of proposed Mizo tagsets and each tag's frequency in the corpus

Tags	**Descriptions**	**Frequency**	**Tag**	**Description**	**Frequency**
VB	Verb base form	11644	MP	Demonstrative Pronoun	332
CMN	Common Noun	8653	DJJ	Double Adjective	322
RB	Adverb base form	8076	;	Semi colon	279
PSP	Personal Pronoun	7740	RBM	Adverb of Manner	203
PPN	Proper Noun	6471	(	Opening bracket	184
FW	Foreign Word	4214	)	Closing bracket	181
,	Comma	3301	VBN	Verbal Noun	181
PT	Particles	2740	POP	Possessive Pronoun	172
CC	Coordinating Conjunction	2699	RLP	Relative Pronoun	149
JJ	Adjective base form	2637	SJJ	Superlative Adjective	106
.	Full stop	2469	-	Hyphen	100
PPT	Postposition	2065	SRB	Superlative Adverb	95
RBP	Adverb of Place	1886	SF	Suffix	81
AT	Article	1725	NVB	Nounal Verb	62
MJJ	Demonstrative Adjective	1429	CJJ	Comparative Adjective	60
CD	Cardinal Number	1177	UH	Interjection	50
NG	Negation	1158	IJJ	Interrogative Adjective	50
ABN	Abstract Noun	1148	SYM	Symbol	49
MRB	Demonstrative Adverb	935	IRB	Interrogative Adverb	41
RBT	Adverb of Time	776	:	Colon	40
SPRB	Specifying Adverb	565	CRB	Comparative Adverb	37
QM	Quotation Mark	560	DVB	Double Verb	30
DRB	Double Adverb	475	IP	Interrogative Pronoun	29
ET	Date	404	?	Question mark	25

each fold serving as a test set once. The results were then averaged to provide a comprehensive comparison.

6.6.2 Result Analysis and Observations

In this experiment, accuracy serves as the fundamental metric for evaluating the taggers' performance. It is conceptualized as the ratio of words that have been correctly tagged to the overall count of words present in the dataset. Figure 6.3 illustrates a labeled bar chart that displays the average performance of different backoff N-gram taggers. These include the unigram backoff tagger with and without the regular expression (RE) tagger, the bigram backoff tagger with and without the regular expression (RE) tagger, and the trigram backoff tagger with and without the regular expression (RE) tagger.

The findings reveal that integrating the proposed regular expression (RE) tagger with the N-gram backoff taggers results in enhanced accuracy for all three taggers. Specifically, the accuracy increases from 78.68% to 79.88% for the unigram tagger. The bigram tagger shows an accuracy increase from 80.96% to 83.19%, while the trigram tagger exhibits an accuracy improvement from 80.34% to 82.59%. The integration of the regular expression tagger with the N-gram taggers consistently enhances the performance of all taggers, resulting in an average accuracy improvement of 1.89%.

Among the various taggers evaluated, the bigram backoff tagger combined with the regular expression tagger achieved the highest accuracy rate of 83.19%. However, it is noteworthy that the trigram backoff tagger closely competed with the bigram tagger in terms of accuracy. In some cases, after multiple trials, the trigram backoff tagger even outperformed the bigram backoff tagger, indicating that the choice between these two taggers may depend on the specific context or dataset being used.

Based on the results of our experiment, it was observed that the NLTK bigram

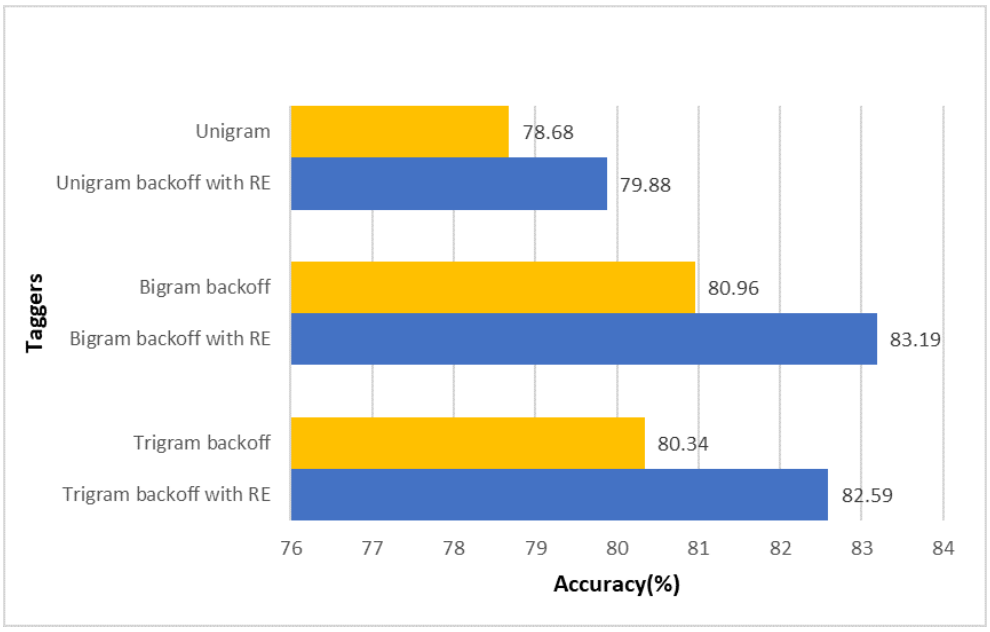

Figure 6.3: N-gram results comparison

tagger outperformed the NLTK bigram tagger by a slight margin. Therefore, we decided to incorporate the bigram tagger into our proposed hybrid approach, which combined an HMM-based method and a rule-based tagger.

The accuracies obtained from the second phase experiment are presented in Table 6.3. The table showcases the performance of three different taggers: the baseline HMM tagger, the first proposed method (where we made modifications to the Viterbi algorithm), and the second proposed method (a hybrid tagger that combines the HMM tagger, N-gram tagger, and rule-based tagger). The baseline Hidden Markov Model (HMM) tagger produced an accuracy of 84.18%. The accuracy increased to 85.32% with the introduction of the first proposed approach and further improved to 89.53% with the hybrid tagger. These results indicate that the first proposed method improved the baseline HMM tagger by approximately 1.14%, and the hybrid approach further enhanced it by around 4.21%. This demonstrates the effectiveness of the proposed techniques in enhancing the accuracy of the HMM-based taggers.

Table 6.3: Performance of HMM-based taggers

Name of Taggers	**Accuracy**
Baseline HMM tagger	84.18%
Proposed method I	85.32%
Proposed hybrid tagger	89.53%

6.6.3 Detail Evaluation on The Proposed Hybrid Tagger

We conducted a comprehensive analysis of the performance of our newly devised hybrid POS tagger, which stands as the ultimate approach and attains the highest accuracy score. Our meticulous examination of the hybrid POS tagger's effectiveness encompassed the assessment of precision, recall, and F1-score for each individual tag (detailed definitions of these metrics were previously elaborated upon in the preceding chapter). The outcomes of this analysis are highlighted in Table 6.5, revealing promising and encouraging results.

Moreover, to provide a comprehensive overview, we computed both the macro average and weighted average for these performance metrics. Here, the term "macro average" refers to the mean of the metric scores across all tags, treating each tag equally regardless of its frequency or support. On the other hand, the "weighted average" considers the overall performance while accounting for the different frequencies of tags, providing a more representative summary by giving more weight to tags with higher occurrence rates. These comprehensive summaries are presented in Table 6.4.

Table 6.4: Macro average and Weighted average

	Precision	**Recall**	**F1-score**
Macro avg	86.55%	82.37%	82.41%
Weighted avg	85.14%	84.83%	84.48%

The macro average precision, recall, and F1-score were determined to be 86.55.0%, 82.37%, and 82.41%, respectively. These values represent the average performance across all tags without considering the class imbalance. By considering each tag equally, the macro average provides insights into the overall

Table 6.5: Precision, recall, and F1-score for each tag

Tags	Precision	Recall	F1-score	Support	Tags	Precision	Recall	F1-score	Support
-	1.00	0.63	0.77	0.0013	MJJ	0.87	0.85	0.86	0.0184
(	1.00	1.00	1.00	0.0023	MP	0.84	0.95	0.89	0.0043
)	0.60	1.00	0.90	0.0023	MRB	0.91	0.79	0.85	0.0120
,	1.00	0.99	0.99	0.0424	NG	0.97	0.86	0.97	0.0149
.	1.00	1.00	1.00	0.0317	NVB	0.47	0.70	0.56	0.0008
:	1.00	1.00	1.00	0.0005	POP	0.94	0.47	0.63	0.0022
;	1.00	1.00	1.00	0.0036	PPN	0.88	0.94	0.91	0.0831
?	1.00	1.00	1.00	0.0003	PPT	0.80	0.94	0.85	0.0265
ABN	0.95	0.90	0.90	0.0147	PSP	0.90	0.78	0.84	0.0994
AT	0.67	0.74	0.70	0.0222	PT	0.79	0.90	0.84	0.0352
CC	0.92	0.75	0.83	0.0347	QM	1.00	0.91	0.95	0.0072
CD	0.89	1.00	0.94	0.0151	RB	0.70	0.89	0.77	0.1038
CJJ	0.96	0.89	0.92	0.0008	RBM	0.90	0.67	0.77	0.0026
CMN	0.88	0.78	0.83	0.1112	RBP	0.91	0.72	0.80	0.0242
CRB	0.90	0.72	0.80	0.0005	RBT	0.91	0.63	0.75	0.0100
DJJ	1.00	1.00	1.00	0.0041	RLP	0.93	0.54	0.68	0.0019
DRB	0.96	0.82	0.88	0.0061	SF	0.78	0.46	0.58	0.0010
DVB	0.17	0.91	0.29	0.0004	SJJ	0.90	0.47	0.62	0.0014
ET	1.00	0.33	0.50	0.0052	SPRB	0.83	0.67	0.74	0.0073
FW	0.80	0.86	0.83	0.0541	SRB	1.00	0.67	0.80	0.0012
IJJ	0.92	0.87	0.89	0.0006	SYM	1.00	1.00	1.00	0.0006
IP	0.12	0.56	0.20	0.0004	UH	0.45	0.86	0.59	0.0006
IRB	0.92	0.8	0.86	0.0005	VB	0.70	0.84	0.76	0.1496
JJ	0.87	0.76	0.81	0.0339	VBN	0.67	0.9	0.77	0.0024

effectiveness of the tagger.

Additionally, we calculated the weighted average precision, recall, and F1-score, which were found to be 85.14%, 84.83%, and 84.48%, respectively. The weighted average takes into account the class imbalance issue by considering the contribution of each tag proportional to its occurrence in the dataset. These metrics demonstrate the tagger's ability to handle class imbalance effectively and provide a more accurate representation of its overall performance.

Analyzing the results further, we can conclude that the hybrid POS tagger exhibited a balanced performance across all tags, as indicated by the average macro scores. This suggests that the tagger performed consistently well for various tags without significant variations in performance among them.

Moreover, the weighted average scores reinforce the effectiveness of the tagger in addressing the class imbalance issue. By giving more weight to the underrepresented tags, the tagger achieved good overall performance, showcasing its ability to handle imbalanced datasets effectively.

Based on these findings, we can firmly assert that our hybrid POS tagger demonstrated good performance, with balanced results across all tags and effective handling of class imbalance. These results validate the reliability and accuracy of our tagger, providing a solid foundation for its application in various natural language processing tasks.

6.7 Performance Comparison

This section provides a detailed examination of the performance of the baseline HMM model, proposed method I, and method II (Hybrid tagger). The purpose here is to delve into how well these taggers perform when it comes to accurately assigning labels to words within the designated context. Interestingly, while both taggers managed to accurately label a set of words, there were instances where

their performance fell short. Notably, certain words were mislabeled by both the baseline HMM model and the newly proposed hybrid HMM-based tagger. This suggests that these words posed a challenge for both systems and highlights the complexity of accurately tagging certain linguistic elements. Examples of such instances from the corpus are shown in Table 6.6.

Table 6.6: Performance of baseline HMM, Method I and Method II (Hybrid tagger) on selected words

Words	**Baseline HMM**	**Method I**	**Method II**	**Correct tag**
yellow	None	CMN	FW	FW
inkhelh	None	VB	VB	VB
Lammualah	None	CMN	VB	RBP
Anmahni	None	PPN	PPN	PSP
thutlukna	VBN	VBN	ABN	ABN
Sawrkar	CMN	CMN	VBN	VBN
Sawrkar	CMN	CMN	CMN	CMN
khatah	RBT	RBT	RBT	RBP
Boxer	CMN	CMN	CMN	FW

When the standard HMM tagger encounters a word not found in its vocabulary, it assigns the tag "None" to that word. It has been observed that the performance of the proposed I closely resembles that of the standard HMM tagger, except when dealing with unknown words. Additionally, certain words in the corpus are found to carry multiple tags due to variations in their contextual usage. For example, the term *'sawrkar'* has been labeled as both "CMN" and "VBN" based on the context within the corpus. While the standard HMM tagger and the proposed I tagger consistently tag all instances of this word as "CMN" regardless of context, the proposed hybrid tagger appropriately assigns both "CMN" and "VBN" tags based on the relevant contextual cues.

Let's consider another instance, the word 'boxer.' As per the rules specified in the proposed tagger, *'boxer'* would have received an FW label. However, because the hybrid tagger determines the most likely tag using probability methods, it mistakenly tags it as "CMN", similar to the other two taggers.

Figures 6.4, 6.5 and 6.6 display bar graphs illustrating the occurrences of erroneous predictions by the baseline HMM tagger, Method 1, and the suggested hybrid tagger (Method 2). Notably, the baseline and Method 1 exhibit a nearly equivalent frequency of incorrect predictions, both of which are notably higher in comparison to the erroneous predictions made by the proposed hybrid tagger. Examining the trends depicted in 6.4, and 6.5, it becomes apparent that the error rate is significantly impacted by the RBP, which has been notably reduced in Figure 6.6. Moreover, the prediction errors in Figure 6.6 are evenly distributed.

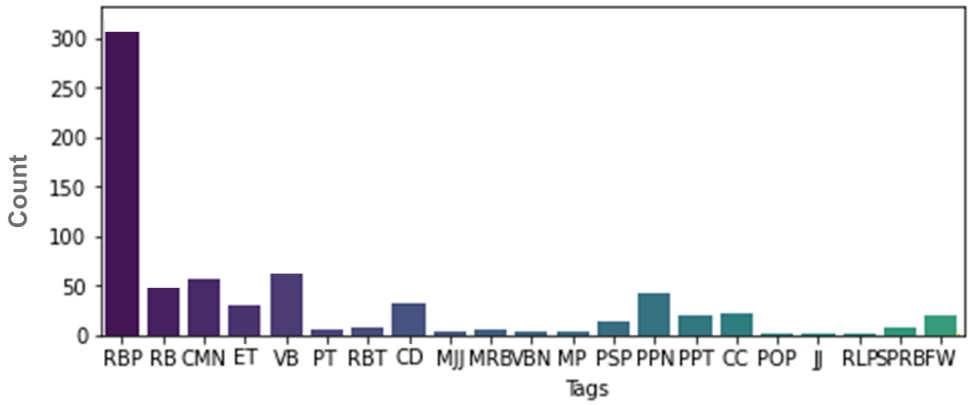

Figure 6.4: No. of incorrect predictions for baseline HMM tagger

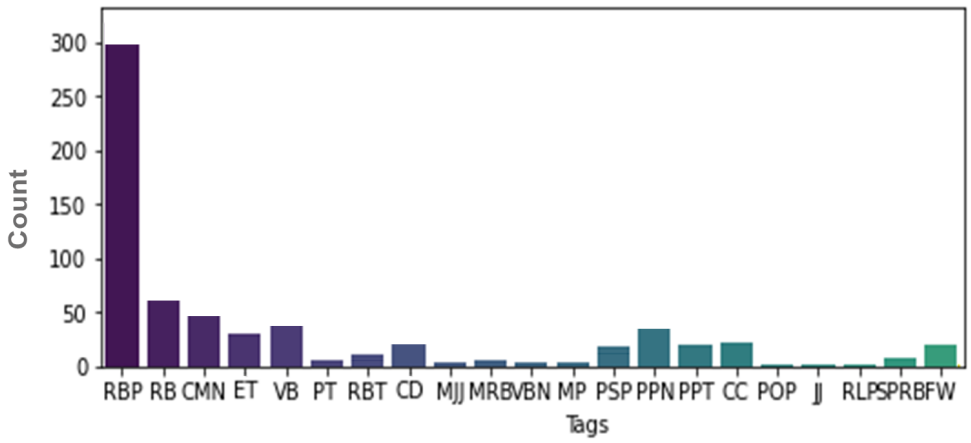

Figure 6.5: No. incorrect predictions for Proposed 1

We also compare the two taggers' performance using external test data. Three simple sentences containing known and unknown words are chosen and input into

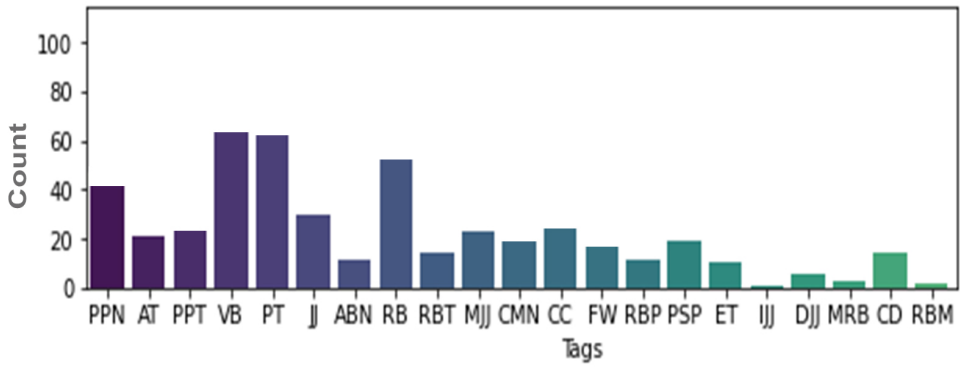

Figure 6.6: No. incorrect predictions for the proposed hybrid tagger

the system to examine how the taggers assign tags to each token.

Input text 1 : *"Aizawlah ka bazar dawn"*

Input text 2 : *"Dintharah ka hau dawn"*

The above two simple sentences are handled quite well by the two taggers as shown in Table 6.7. In the training corpus, the word *bazar* has been assigned three different tags, such as CMN, VBN, and PPN. When this sample input text 1 is fed to the system, the baseline HMM and our proposed tagger successfully identify the word *bazar* as VBN. The term *hau*, which appears in input text 2 is an unknown word. The baseline HMM tagger wrongly tags the word as "ABN," whereas the proposed tagger correctly tags it as VB.

Input text 3: *"Bombay-ah Mawia remhriatna avangin WPO buatsaih X-ray chungchang seminar-ah ka tel ve."*

The third statement in the input text is a little more complicated than the previous two input samples. Table 6.8 provides a detailed analysis.

The proposed hybrid tagger operates well on this particular sample input. Except for the word *seminar-ah*, which has an incorrect RBP tag, the proposed

Table 6.7: Performance on external test sentences

Input	Taggers	Tagging
Input text 1	Baseline HMM tagger	Aizawlah/RBP ka/PSP bazar/VBN dawn/RB
	Proposed hybrid tagger	Aizawlah/RBP ka/PSP bazar/VBN dawn/RB
Input text 2	Baseline HMM tagger	Dintharah/RBP ka/PSP hau/ABN dawn/RB
	Proposed hybrid tagger	Dintharah/RBP ka/PSP hau/VB dawn/RB

Table 6.8: Performance on external test sentence 2 2

Input	**Known/ Unknown**	**Baseline HMM tagger**	**Proposed hybrid tagger**	**Correct tag**
Bombay-ah	Unknown	ABN	RBP	RBP
Mawia	Unknown	ABN	PPN	PPN
remhriatna	Unknown	ABN	ABN	ABN
avangin	Known	CC	CC	CC
WPO	Unknown	PPN	PPN	PPN
buatsaih	Known	VB	VB	VB
X-ray	Unknown	ABN	FW	FW
chungchang	Known	PPT	PPT	PPT
seminar-ah	Unknown	ABN	RBT	RBP
ka	Known	PSP	PSP	PSP
tel	Known	VB	VB	VB
ve	Known	RB	RB	RB

method appropriately assigned tags to each word. With the help of the rule-based tagger coupled with the proposed system, unknown terms are correctly handled.

6.7.1 Advantages of the Proposed Hybrid Tagger

Upon analyzing the experimental outcomes, several advantages of the hybrid tagger become evident. Some of these are outlined below:

Enhanced robustness and accuracy: The proposed hybrid model excels in capturing nuanced linguistic patterns by amalgamating probabilistic modeling

and rule-based heuristics. This amalgamation equips the model to handle a wide spectrum of linguistic scenarios, making it highly accurate and robust in real-world applications. Extensive experiments on diverse datasets reveal that our hybrid model consistently outperforms the individual tagging methods in terms of precision, recall, and F1 score.

Handling unknown words: When faced with out-of-vocabulary terms, our model seamlessly delegates tagging responsibilities to the rule-based tagger, which employs linguistic heuristics and domain-specific knowledge to make informed tagging decisions.

Scalability and adaptability: Our model exhibits a remarkable level of scalability and adaptability to other languages. This adaptability is achieved by crafting language-specific rules that are seamlessly integrated into the rule-based tagger component.

6.7.2 Challenges with the Proposed Hybrid Tagger

They proposed hybrid model leverages the unique strengths of each tagging method to address various linguistic complexities. However, it is not exempt from encountering its own set of challenges. Some of these challenges are highlighted below:

Limitations of the proposed designed rules: The language rules implemented in this study, while providing valuable heuristics for handling unknown words and linguistic patterns, may require further refinement and updates. The present set of rules may not encompass all possible linguistic variations, adapt swiftly to evolving language patterns, or address domain-specific challenges

Threshold selection: One of the primary challenges we grapple with is threshold selection. While dynamic thresholding is a powerful feature, determining the optimal threshold value requires careful experimentation and validation. Striking the right balance between precision and recall remains an ongoing chal-

lenge.

Computational complexity: Integrating multiple tagging methods introduces a degree of computational complexity. It is imperative to implement optimization strategies and consider parallel processing techniques to maintain efficiency.

6.8 Conclusion

This study presents three distinct methods aimed at enhancing POS tagging in the Mizo language. The initial experiment focuses on improving the N-gram backoff taggers by incorporating a regular expression tagger. Through this approach, we were able to enhance the performance of the N-gram taggers, resulting in an average accuracy improvement of 1.89%.

Two additional methods were proposed to further enhance the baseline HMM model's performance. The first proposed model achieved an accuracy of 85.32%, surpassing the baseline HMM model. The second proposed hybrid tagger demonstrated the highest accuracy of 89.53%, which is 5.35% higher than the baseline HMM model. These results highlight the effectiveness of our approaches in improving the accuracy and overall performance of the tagger for both known and unknown words.

In addition to the experimental work, we conducted a thorough analysis of the Mizo language and its grammatical structure. This analysis provided valuable insights into the morphological clues that can aid in accurately assigning tags to words. By leveraging these insights, our methods were able to substantially improve the accuracy of the tagger.

Considering the low-resource nature of the Mizo language, the contributions and results achieved through our proposed approaches are noteworthy. We believe that further improvements can be made by increasing the corpus size and

designing more precise regular expressions tailored to the language's specific characteristics. In future work, we plan to tackle model complexity by delving into potential strategies. This will involve an examination of parallelization methods, utilizing the capabilities of multi-core processors and distributed computing environments to optimize computational performance. We also to explore using different deep-learning architectures to enhance the accuracy of the Mizo tagger.

Overall, this study serves as a solid foundation for subsequent research in the field, and we anticipate that our efforts will contribute significantly to the advancement of POS tagging in the Mizo language.

Chapter 7

Conclusion AND Future Scope

In our research work, we have explored the different methods of part-of-speech tagging in the context of the Mizo language. The primary goal of this research was to develop effective taggers that can accurately assign part-of-speech tags to words in Mizo sentences, thereby aiding various natural language processing tasks. Throughout the course of this study, we have investigated different approaches and techniques, including Hidden Markov Models (HMM), Conditional Random Fields (CRF), and a Hybrid model that combines the strengths of multiple taggers.

At the outset, we initiated our research by presenting an overview of the research problem and establishing the framework for the subsequent chapters. The journey began by introducing the concept of part-of-speech tagging, emphasizing its significance and wide-ranging applications in various linguistic tasks. This introductory phase laid a solid foundation for our study, underscoring the crucial role of POS tagging in the field of natural language processing and its specific relevance to the Mizo language. We then delved into the existing literature to understand the state-of-the-art techniques and approaches employed in POS tagging. The second chapter presented a comprehensive literature survey, offering a review of relevant studies and existing state-of-the-art techniques and approaches in the field of part-of-speech tagging.

Chapter 3 focused specifically on the Mizo language, delving into its unique characteristics, grammatical features, and challenges associated with POS tagging. Understanding the intricacies of the Mizo language was crucial in developing effective POS tagging models tailored to its specific requirements.

In Chapter 4, we introduced a hidden Markov model (HMM)-based POS tagger specifically designed for the Mizo language. The model was trained on annotated Mizo corpora, enabling it to assign appropriate part-of-speech tags to words based on their contextual information. We discussed the design and implementation details of the proposed model, including the training and evaluation process. Through experimentation, we demonstrated the effectiveness of the HMM-based approach in accurately tagging POS labels in Mizo sentences. The experimental results showed that the HMM-based tagger achieved promising accuracy levels, showcasing its ability to capture the sequential dependencies of words in Mizo sentences.

In Chapter 5, we explored the conditional random fields (CRF) algorithm as an alternative method for POS tagging. We explored the underlying principles of CRFs and their ability to capture complex dependencies between adjacent words, leading to improved tagging accuracy. CRF models provided a more sophisticated approach to capturing contextual information, considering not only the current word but also a wider context of words. The evaluation results indicated that the CRF-based POS tagger achieved competitive performance, suggesting its potential for POS tagging in the Mizo language.

Chapter 6 introduced the Hybrid model, aiming to leverage the strengths of multiple taggers to enhance overall performance. Our goal was to explore innovative and alternative techniques, combining them to overcome the limitations of popular approaches such as HMM-based, N-gram, and rule-based models. By integrating the HMM tagger with the NLTK N-gram tagger and the Rule-based tagger, we created a robust system capable of handling a wider range of linguistic patterns and achieving higher accuracy levels. The experimental evaluation demonstrated the superiority of the Hybrid model, showcasing its ability to outperform the individual three taggers and provide a more comprehensive solution for part-of-speech tagging in Mizo.

7.1 Summary of Our Research Contributions

Our research has made significant contributions to the field of part-of-speech tagging for the Mizo language, unveiling valuable insights into its linguistic peculiarities, grammatical structures, and syntactic patterns. These contributions have paved the way for the development of more accurate and effective part-of-speech taggers specifically tailored for Mizo. The main contributions of the thesis can be summarized as follows:

Development of Mizo tagset: To ensure precise and consistent part-of-speech tagging, we created a specialized Mizo Tagset consisting of 48 tags. This Tagset considers the unique grammatical features and syntactic variations of the Mizo language, enabling more accurate and detailed annotation. The development of this tagset serves as a valuable resource for future research and applications in Mizo language processing.

Creation of tagged corpus: In order to train and evaluate our part-of-speech taggers, we created a comprehensive tagged corpus consisting of 77,027 words. Each word in the corpus was manually annotated with its corresponding part-of-speech tag based on the developed Mizo Tagset. This corpus serves as a valuable resource for training and testing part-of-speech taggers in Mizo, enabling researchers and practitioners to assess the performance and effectiveness of different models and techniques. The availability of such a corpus not only supports further research in part-of-speech tagging but also opens avenues for other natural language processing tasks, including parsing, machine translation, and information extraction.

Insights into linguistic peculiarities: Furthermore, our research has shed light on the unique linguistic characteristics of Mizo. By delving into the language's syntactic structures and grammatical rules, we have gained valuable insights into its underlying patterns and intricacies. This understanding has informed the design and development of our part-of-speech taggers, ensuring that

they capture and leverage the specific features of Mizo to improve their accuracy and performance.

Development of taggers: We have explored and developed various taggers, including HMM, CRF, and a Hybrid model, to address the challenges posed by the linguistic characteristics of Mizo. The experimental evaluations have showcased the effectiveness of these taggers, highlighting their ability to achieve high levels of accuracy in part-of-speech tagging. By developing effective taggers and exploring innovative approaches, we have taken significant steps toward improving the accuracy and performance of natural language processing tasks in the Mizo language domain.

In conclusion, our research has made significant achievements in the field of part-of-speech tagging for the Mizo language. These contributions serve as a foundation for further research and development in part-of-speech tagging for the Mizo language. We anticipate that our research will inspire further investigations into Mizo language processing, contributing to the development of more advanced language technologies that cater to the specific needs of Mizo speakers and researchers.

7.2 Future Research Directions

Although this thesis has offered valuable insights into POS tagging for the Mizo language, there are numerous possibilities for future research and development. Exploring these avenues will lead to improved accuracy and applicability of POS tagging models in the context of the Mizo language. The following points present potential areas for investigation:

Expanding the training data: To improve the performance of POS taggers, it is essential to increase the size and diversity of the training data. Collecting more annotated corpora specifically for the Mizo language will enable the development of more accurate models. This expansion could include texts from various

domains, genres, and sources to ensure robustness and coverage across different linguistic contexts.

Domain-specific POS tagging: Investigating domain-specific POS tagging can enhance the applicability of tagging models in specific fields such as healthcare, finance, or legal domains. Adapting the models to specialized language use can improve their accuracy and utility in real-world applications. Developing domain-specific annotated corpora and incorporating domain-specific linguistic patterns and terminologies will contribute to more precise and effective POS tagging in domain-specific contexts.

Deep learning approaches: Exploring deep learning techniques, such as recurrent neural networks (RNNs) or transformers, holds promise for advancing POS tagging in the Mizo language. These models have shown impressive performance in various natural language processing tasks and can potentially capture complex linguistic patterns and dependencies. Integrating deep learning architectures into POS tagging systems for Mizo could improve overall accuracy and handle the language's specific challenges.

Designing more precise rules for the Mizo language: Building upon the rule-based tagger discussed in Chapter 6, further research can focus on designing more precise linguistic rules that cater specifically to the intricacies of the Mizo language. Analyzing the language's syntactic structures, morphological variations, and semantic nuances will enable the development of rule sets that capture the unique characteristics of Mizo. Such rules can complement statistical models and contribute to fine-grained POS tagging in Mizo.

By pursuing these future directions, researchers can make further advancements in the field of POS tagging in the Mizo language and contribute to the development of more accurate, robust, and linguistically-informed tagging models. These endeavors will not only benefit natural language processing applications specific to Mizo but also shed light on the challenges and opportunities in processing other morphologically rich and low-resource languages.

REFERENCES

[1] D. Swade and C. Babbage, *Difference engine: Charles Babbage and the quest to build the First Computer*. Viking Penguin, 2001.

[2] E. F. Koerner, *Ferdinand de Saussure: Origin and development of his linguistic thought in western studies of language*, vol. 7. Springer-Verlag, 2013.

[3] F. De Saussure, *Course in general linguistics*. Columbia University Press, 2011.

[4] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition."

[5] U. Tiwary and T. Siddiqui, *Natural language processing and information retrieval*. Oxford University Press, Inc., 2008.

[6] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[7] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis lectures on human language technologies*, vol. 10, no. 1, pp. 1–309, 2017.

[8] A. A. Patel and A. U. Arasanipalai, *Applied Natural Language Processing in the Enterprise*. " O'Reilly Media, Inc.", 2021.

[9] C. Dos Santos and B. Zadrozny, "Learning character-level representations for part-of-speech tagging," in *International Conference on Machine Learning*, pp. 1818–1826, PMLR, 2014.

[10] C. D. Manning, "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?," in *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I 12*, pp. 171–189, Springer, 2011.

[11] M. Bano, "Addressing the challenges of requirements ambiguity: A review of empirical literature," in *2015 IEEE Fifth International Workshop on Empirical Requirements Engineering (EmpiRE)*, pp. 21–24, IEEE, 2015.

[12] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Third conference on applied natural language processing*, pp. 133–140, 1992.

[13] M. Sun and J. R. Bellegarda, "Improved pos tagging for text-to-speech synthesis," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5384–5387, IEEE, 2011.

[14] L. R. Teodorescu, R. Boldizsar, M. Ordean, M. Duma, L. Detesan, and M. Ordean, "Part of speech tagging for romanian text-to-speech system," in *2011 13th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pp. 153–159, IEEE, 2011.

[15] G. I. Schlünz, *The effects of part–of–speech tagging on text–to–speech synthesis for resource–scarce languages*. PhD thesis, North-West University, 2010.

[16] P. Ganesh, B. S. Rawal, A. Peter, and A. Giri, "Pos-tagging based neural machine translation system for european languages using transformers," *WSEAS Transactions on Information Science and Applications*, vol. 18, pp. 26–33, 2021.

[17] X. Feng, Z. Feng, W. Zhao, B. Qin, and T. Liu, "Enhanced neural machine translation by joint decoding with word and pos-tagging sequences," *Mobile Networks and Applications*, vol. 25, no. 5, pp. 1722–1728, 2020.

[18] P. Alva and V. Hegde, "Hidden markov model for pos tagging in word sense disambiguation," in *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 279–284, IEEE, 2016.

[19] D. McCarthy, "Word sense disambiguation: An overview," *Language and Linguistics compass*, vol. 3, no. 2, pp. 537–558, 2009.

[20] P. Resnik, "A perspective on word sense disambiguation methods and their evaluation," in *Tagging Text with Lexical Semantics: Why, What, and How?*, 1997.

[21] P. A. Heeman and J. F. Allen, "Incorporating pos tagging into language modeling," *arXiv preprint cmp-lg/9705014*, 1997.

[22] M. Nakamura, K. Maruyama, T. Kawabata, and K. Shikano, "Neural network approach to word category prediction for english texts," in *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*, 1990.

[23] G. Kanaan, R. Al-Shalabi, and M. Sawalha, "Improving arabic information retrieval systems using part of speech tagging," *Information Technology Journal*, vol. 4, no. 1, pp. 32–37, 2005.

[24] I. Gashaw and H. Shashirekha, "Enhanced amharic-arabic cross-language information retrieval system using part of speech tagging," in *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pp. 1–7, IEEE, 2019.

[25] R. Karimpour, A. Ghorbani, A. Pishdad, M. Mohtarami, A. AleAhmad, H. Amiri, and F. Oroumchian, "Using part of speech tagging in persian information retrieval.," in *CLEF (Working Notes)*, Citeseer, 2008.

[26] H. Bast and E. Haussmann, "More accurate question answering on freebase," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, pp. 1431–1440, 2015.

[27] Z. Z. Hlaing, Y. K. Thu, T. Supnithi, and P. Netisopakul, "Improving neural machine translation with pos-tag features for low-resource language pairs," *Heliyon*, vol. 8, no. 8, p. e10375, 2022.

[28] R. Sennrich and B. Haddow, "Linguistic input features improve neural machine translation," *arXiv preprint arXiv:1606.02892*, 2016.

[29] A. Chowdhury and M. C. McCabe, "Improving information retrieval systems using part of speech tagging," tech. rep., 1998.

[30] R. Karimpour, A. Ghorbani, A. Pishdad, M. Mohtarami, A. AleAhmad, H. Amiri, and F. Oroumchian, "Improving persian information retrieval systems using stemming and part of speech tagging," in *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 89–96, Springer, 2008.

[31] S. Rathod and S. Govilkar, "Survey of various pos tagging techniques for indian regional languages," *Int. J. Comput. Sci. Inf. Technol*, vol. 6, no. 3, pp. 2525–2529, 2015.

[32] E. Brill, "A simple rule-based part of speech tagger," tech. rep., Pennsylvania Univ Philadelphia Dept of Computer and Information Science, 1992.

[33] C. Aone and K. Hausman, "Unsupervised learning of a rule-based spanish part of speech tagger," in *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

[34] K. Purnamasari and I. Suwardi, "Rule-based part of speech tagger for indonesian language," in *IOP Conference Series: Materials Science and Engineering*, vol. 407, p. 012151, IOP Publishing, 2018.

[35] B. Pham, "Parts of speech tagging: Rule-based," 2020.

[36] P. K. Vaishali, K. Kalpana, and C. Namrata Mahender, "A rule-based approach for marathi part-of-speech tagging," in *ICT with Intelligent Applications: Proceedings of ICTIS 2021, Volume 1*, pp. 773–785, Springer, 2022.

[37] W. S. Stolz, P. H. Tannenbaum, and F. V. Carstensen, "Stochastic approach to the grammatical coding of english," *Communications of the ACM*, vol. 8, no. 6, pp. 399–405, 1965.

[38] J. Kupiec, "Robust part-of-speech tagging using a hidden markov model," *Computer speech & language*, vol. 6, no. 3, pp. 225–242, 1992.

[39] S. K. Sharma and G. S. Lehal, "Using hidden markov model to improve the accuracy of punjabi pos tagger," in *2011 IEEE International Conference on Computer Science and Automation Engineering*, vol. 2, pp. 697–701, IEEE, 2011.

[40] C. D. E. Reyes, K. R. S. Suba, A. R. Razon, and P. C. Naval Jr, "Svpost: A part-of-speech tagger for tagalog using support vector machines," in *Proceedings of the 11th Philippine Computing Science Congress*, 2011.

[41] M. Murata, Q. Ma, and H. Isahara, "Part of speech tagging in thai language using support vector machine," *arXiv preprint cs/0112004*, 2001.

[42] H. Schmid, "Probabilistic part-ofispeech tagging using decision trees," in *New methods in language processing*, p. 154, 2013.

[43] L. Màrquez and H. Rodríguez, "Part-of-speech tagging using decision trees," in *European conference on machine learning*, pp. 25–36, Springer, 1998.

[44] A. Dalal, K. Nagaraj, U. Sawant, and S. Shelke, "Hindi part-of-speech tagging and chunking: A maximum entropy approach," *Proceeding of the NLPAI Machine Learning Competition*, 2006.

[45] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Maximum entropy based bengali part of speech tagging," *A. Gelbukh (Ed.), Advances in Natural Language Processing and Applications, Research in Computing Science (RCS) Journal*, vol. 33, pp. 67–78, 2008.

[46] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[47] A. Swamy and S. Srinath, "Pos tagging and ner system for kannada using conditional random fields," *International Journal of Information Retrieval Research (IJIRR)*, vol. 11, no. 4, pp. 1–13, 2021.

[48] S. Warjri, P. Pakray, S. A. Lyngdoh, and A. K. Maji, "Part-of-speech (pos) tagging using conditional random field (crf) model for khasi corpora," *International Journal of Speech Technology*, vol. 24, no. 4, pp. 853–864, 2021.

[49] W. Khan, A. Daud, J. A. Nasir, T. Amjad, S. Arafat, N. Aljohani, and F. S. Alotaibi, "Urdu part of speech tagging using conditional random fields," *Language Resources and Evaluation*, vol. 53, pp. 331–362, 2019.

[50] A. Ajees and S. M. Idicula, "A pos tagger for malayalam using conditional random fields," *Int. J. Appl. Eng. Res*, vol. 13, no. 3, 2018.

[51] M. Dibitso, P. A. Owolawi, and S. O. Ojo, "An hybrid part of speech tagger for setswana language using a voting method," in *International Conference on Intelligent and Innovative Computing Applications*, pp. 245–253, 2022.

[52] C. Tailor and B. Patel, "Hybrid pos tagger for gujarati text," in *Soft Computing and its Engineering Applications: Second International Conference, icSoftComp 2020, Changa, Anand, India, December 11–12, 2020, Proceedings 2*, pp. 134–144, Springer, 2021.

[53] C. Kruengkrai, K. Uchimoto, Y. Wang, K. Torisawa, H. Isahara, *et al.*, "An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 513–521, 2009.

[54] R. Simionescu, "Hybrid pos tagger," in *Proceedings of Language Resources and Tools with Industrial Applications Workshop (Eurolan 2011 Summer School), Cluj-Napoca, Romania*, pp. 21–28, Citeseer, 2011.

[55] H. Schmid, "Part-of-speech tagging with neural networks," *arXiv preprint cmp-lg/9410018*, 1994.

[56] J. A. Ovi, M. A. Islam, and M. R. Karim, "Banep: An end-to-end neural network based model for bangla parts-of-speech tagging," *IEEE Access*, vol. 10, pp. 102753–102769, 2022.

[57] A. Tehseen, T. Ehsan, H. B. Liaqat, A. Ali, and A. Al-Fuqaha, "Neural pos tagging of shahmukhi by using contextualized word representations," *Journal of King Saud University-Computer and Information Sciences*, vol. 35, no. 1, pp. 335–356, 2023.

[58] Y. An, X. Xia, X. Chen, F.-X. Wu, and J. Wang, "Chinese clinical named entity recognition via multi-head self-attention based bilstm-crf," *Artificial Intelligence in Medicine*, vol. 127, p. 102282, 2022.

[59] X. Yu, A. Faleńska, and N. T. Vu, "A general-purpose tagger with convolutional neural networks," *arXiv preprint arXiv:1706.01723*, 2017.

[60] M. Otman, B. Mohamed, *et al.*, "Amazigh part of speech tagging using gated recurrent units (gru)," in *2021 7th International Conference on Optimization and Applications (ICOA)*, pp. 1–6, IEEE, 2021.

[61] R. Saidi, F. Jarray, and M. Mansour, "A bert based approach for arabic pos tagging," in *Advances in Computational Intelligence: 16th International Work-Conference on Artificial Neural Networks, IWANN 2021, Virtual Event, June 16–18, 2021, Proceedings, Part I 16*, pp. 311–321, Springer, 2021.

[62] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," *Journal of Big Data*, vol. 9, no. 1, pp. 1–25, 2022.

[63] United Nations Education, Scientific and Cultural Organisation, "Atlas of the world's languages in danger." `http://www.unesco.org`, Last accessed on 13.9.2022.

[64] W. N. Francis and H. Kucera, "Brown corpus manual," *Letters to the Editor*, vol. 5, no. 2, p. 7, 1979.

[65] M. Marcus, B. Santorini, and M. A. Marcinkiewicz, "Building a large annotated corpus of english: The penn treebank," 1993.

[66] M.-C. De Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal dependencies," *Computational linguistics*, vol. 47, no. 2, pp. 255–308, 2021.

[67] N. Ide, "The american national corpus: Then, now, and tomorrow," in *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, Summerville, MA. Cascadilla Proceedings Project*, 2008.

[68] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," *arXiv preprint arXiv:1104.2086*, 2011.

[69] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "Ontonotes: the 90% solution," in *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pp. 57–60, 2006.

[70] N. Choudhary, "Ldc-il: The indian repository of resources for language technology," *Language Resources and Evaluation*, vol. 55, no. 3, pp. 855–867, 2021.

[71] S. Klein and R. F. Simmons, "A computational approach to grammatical coding of english words," *Journal of the ACM (JACM)*, vol. 10, no. 3, pp. 334–347, 1963.

[72] Z. Harris, "String analysis of language structure," *Mouton and Co., The Hague*, 1962.

[73] L. R. Bahl and R. L. Mercer, "Part of speech assignment by a statistical decision algorithm," in *IEEE International Symposium on Information Theory*, pp. 88–89, 1976.

[74] B. B. Greene and G. M. Rubin, *Automatic grammatical tagging of English.* Department of Linguistics, Brown University, 1971.

[75] F. Karlsson, A. Voutilainen, J. Heikkilae, and A. Anttila, *Constraint Grammar: a language-independent system for parsing unrestricted text*, vol. 4. Walter de Gruyter, 2011.

[76] M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, "Uzbek-tagger: The rule-based pos tagger for uzbek language," *arXiv preprint arXiv:2301.12711*, 2023.

[77] F. Rashel, A. Luthfi, A. Dinakaramani, and R. Manurung, "Building an indonesian rule-based part-of-speech tagger," in *2014 international conference on Asian language processing (IALP)*, pp. 70–73, IEEE, 2014.

[78] M. Alex and L. Q. Zakaria, "Brill's rule-based part of speech tagger for kadazan," *International Journal on Recent Trends in Engineering & Technology*, vol. 10, no. 1, p. 75, 2014.

[79] E. Roche and Y. Schabes, "Deterministic part-of-speech tagging with finite state transducers," *Computational linguistics*, vol. 21, no. 2, pp. 227–253, 1995.

[80] T. Brants, "Tnt-a statistical part-of-speech tagger," *arXiv preprint cs/0003055*, 2000.

[81] S. Dandapat and S. Sarkar, "Part of speech tagging for bengali with hidden markov model," *Proceeding of the NLPAI machine learning competition*, 2006.

[82] S.-Z. Lee, J. Tsujii, and H. C. Rim, "Lexicalized hidden markov models for part-of-speech tagging," in *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics*, 2000.

[83] D. L. Cing and K. M. Soe, "Improving accuracy of part-of-speech (pos) tagging using hidden markov model and morphological analysis for myanmar language," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 2, p. 2023, 2020.

[84] A. F. Huda, M. H. N. Hilal, A. Saepuloh, and D. Supriadi, "Analysis part of speech tagging using hidden markov model on qur'an data," in *2021 7th International Conference on Wireless and Telematics (ICWT)*, pp. 1–6, IEEE, 2021.

[85] S. F. Adafre, "Part of speech tagging for amharic using conditional random fields," in *Proceedings of the ACL workshop on computational approaches to semitic languages*, pp. 47–54, 2005.

[86] F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic part of speech tagging for bahasa indonesia," in *Third international MALINDO workshop*, pp. 1–6, 2009.

[87] M. Silfverberg, T. Ruokolainen, K. Linden, and M. Kurimo, "Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy," *Unknown host publication*, 2014.

[88] A. Fanoon and G. Uwanthika, "Part of speech tagging for twitter conversations using conditional random fields model," in *2019 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 108–112, IEEE, 2019.

[89] H. Huang and X. Zhang, "Part-of-speech tagger based on maximum entropy model," in *2009 2nd IEEE International Conference on Computer Science and Information Technology*, pp. 26–29, IEEE, 2009.

[90] D. E. Cahyani and W. Mustikaningtyas, "Indonesian part of speech tagging using maximum entropy markov model on indonesian manually tagged corpus," *IAES International Journal of Artificial Intelligence*, vol. 11, no. 1, p. 336, 2022.

[91] C. Yi, "An english pos tagging approach based on maximum entropy," in *2015 International Conference on Intelligent Transportation, Big Data and Smart City*, pp. 81–84, IEEE, 2015.

[92] A. Y. Kassahun and T. G. Fantaye, "Design and develop a part of speech tagging for ge'ez language using deep learning approach," in *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pp. 66–71, IEEE, 2022.

[93] T. Dalai, T. K. Mishra, and P. K. Sa, "Part-of-speech tagging of odia language using statistical and deep learning-based approaches," *arXiv preprint arXiv:2207.03256*, 2022.

[94] Q. Xu and Z. Wang, "A data-driven model for automated chinese word segmentation and pos tagging," *Computational Intelligence and Neuroscience: CIN*, vol. 2022, 2022.

[95] C. Lv, H. Liu, Y. Dong, F. Li, and Y. Liang, "Using uniform-design gep for part-of-speech tagging," *Journal of Circuits, Systems and Computers*, vol. 26, no. 04, p. 1750060, 2017.

[96] C. Lv, H. Liu, and Y. Dong, "An efficient corpus based part-of-speech tagging with gep," in *2010 Sixth International Conference on Semantics, Knowledge and Grids*, pp. 289–292, IEEE, 2010.

[97] X. Xue and J. Zhang, "Part-of-speech tagging of building codes empowered by deep learning and transformational rules," *Advanced Engineering Informatics*, vol. 47, p. 101235, 2021.

[98] A. Imani, S. Severini, M. J. Sabet, F. Yvon, and H. Schütze, "Graph-based multilingual label propagation for low-resource part-of-speech tagging," *arXiv preprint arXiv:2210.09840*, 2022.

[99] S. Besharati, H. Veisi, A. Darzi, and S. H. H. Saravani, "A hybrid statistical and deep learning based technique for persian part of speech tagging," *Iran Journal of Computer Science*, vol. 4, pp. 35–43, 2021.

[100] P. Ranjan and H. Basu, "Part of speech tagging and local word grouping techniques for natural language parsing in hindi," in *Proceedings of the 1st International Conference on Natural Language Processing (ICON 2003)*, Citeseer, 2003.

[101] M. Shrivastava, N. Agrawal, S. Singh, and P. Bhattacharya, "Harnessing morphological analysis in pos tagging task," *Proceedings ICON*, 2005.

[102] S. Singh, K. Gupta, M. Shrivastava, and P. Bhattacharyya, "Morphological richness offsets resource demand–experiences in constructing a pos tagger for hindi," in *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp. 779–786, 2006.

[103] K. Sarkar and V. Gayen, "A trigram hmm-based pos tagger for indian languages," in *Proceedings of the international conference on frontiers of intelligent computing: theory and applications (FICTA)*, pp. 205–212, Springer, 2013.

[104] N. Joshi, H. Darbari, I. Mathur, *et al.*, "Hmm based pos tagger for hindi," in *Proceeding of 2013 international conference on artificial intelligence, soft computing (AISC-2013)*, pp. 341–349, 2013.

[105] V. Gupta, N. Joshi, and I. Mathur, "Pos tagger for urdu using stochastic approaches," in *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*, pp. 1–5, 2016.

[106] B. G. Patra, K. Debbarma, D. Das, and S. Bandyopadhyay, "Part of speech (pos) tagger for kokborok," in *Proceedings of COLING 2012: Posters*, pp. 923–932, 2012.

[107] S. K. Bharti, R. K. Gupta, S. Patel, and M. Shah, "Context-based bigram model for pos tagging in hindi: A heuristic approach," *Annals of Data Science*, pp. 1–32, 2022.

[108] R. D. Deshmukh and A. Kiwelekar, "Deep learning techniques for part of speech tagging by natural language processing," in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp. 76–81, IEEE, 2020.

[109] M. Rajani Shree and B. Shambhavi, "Pos tagger model for south indian language using a deep learning approach," in *ICCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering*, pp. 155–167, Springer, 2022.

[110] K. Akhil, R. Rajimol, and V. Anoop, "Parts-of-speech tagging for malayalam using deep learning techniques," *International Journal of Information Technology*, vol. 12, pp. 741–748, 2020.

[111] V. Advaith, A. Shivkumar, and B. S. Lakshmi, "Parts of speech tagging for kannada and hindi languages using ml and dl models," in *2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1–5, IEEE, 2022.

[112] R. K. Mundotiya, V. Kumar, A. Mehta, and A. K. Singh, "Attention-based domain adaption using transfer learning for part-of-speech tagging: an experiment on the hindi language," in *Proceedings of the 34th Pacific Asia conference on language, information and computation*, pp. 471–477, 2020.

[113] A. Rajan, A. Salgaonkar, and A. Shaikh, "Deep learning for part of speech (pos) tagging: Konkani," in *Smart Trends in Computing and Communications: Proceedings of SmartCom 2022*, pp. 337–346, Springer, 2022.

[114] M. Sathsarani, T. Thalawaththa, N. Galappaththi, J. Danthanarayana, and A. Gamage, "Sinhala part of speech tagger using deep learning techniques," in *2022 6th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pp. 1–6, IEEE, 2022.

[115] S. Anbukkarasi and S. Varadhaganapathy, "Deep learning based tamil parts of speech (pos) tagger," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 2021.

[116] D. Dutta, S. Halder, and T. Gayen, "Intelligent part of speech tagger for hindi," *Procedia Computer Science*, vol. 218, pp. 604–611, 2023.

[117] P. Pakray, A. Pal, G. Majumder, and A. Gelbukh, "Resource building and parts-of-speech (pos) tagging for the mizo language," in *2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)*, pp. 3–7, IEEE, 2015.

[118] Ethnologue, Languages of the World, "Ethnologue, 25th edition." `https://www.ethnologue.com/ethnoblog/gary-simons/welcome-25th-edition`, Last accessed on 12.2.2023.

[119] B. Lalthangliana, *Culture and folklore of Mizoram*. Publications Division Ministry of Information & Broadcasting, 2005.

[120] J. H. Lorrain and F. W. Savidge, *A grammar and dictionary of the Lushai language (Dulien dialect)*. Assam Secretariat Print. Office, 1898.

[121] L. Khiangte, "The milieu," *Northeast India: A Reader*, 2018.

[122] T. H. Lewin, *Progressive Colloquial Exercises in the Lushai Dialect of the Dzo Or Kúki Language, with Vocabularies and Popular Tales (notated) Thomas Herbert Lewin*. Calcutta central Press Company, limited, 1874.

[123] L. Renthlei, "A study of selected literary translations: English–mizo," 2018.

[124] K. Zawla, *Mizo Tawng Grammar*. K. Zawla, 1969.

[125] Remkunga, *Mizo Tawng Grammar Thar*. Remkunga, 1978.

[126] K. Lalzarzova, "Mizo tawng grammar & composition," *K. Sangzawna, Aizawl, Mizoram*, 2016.

[127] P. Thangzikpuia, "Mizo tawng grammar(based on its usage and unique features)," 2019.

[128] R. Lalhluna, "Cinque foils – zo awng grammar," 2014.

[129] L. T. Fanai, "Tones in mizo language," *Journal of Humanities Social Sciences*, vol. 1, no. 1, 2015.

[130] R. Zothanliana, *'A Study of the Development of Mizo Language in Relation to Word Formation*. PhD thesis, Mizoram University, 2020.

[131] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[132] B. Mor, S. Garhwal, and A. Kumar, "A systematic review of hidden markov models and their applications," *Archives of computational methods in engineering*, vol. 28, pp. 1429–1448, 2021.

[133] F. Al Shamsi and A. Guessoum, "A hidden markov model-based pos tagger for arabic," in *Proceeding of the 8th international conference on the statistical analysis of textual data, France*, pp. 31–42, 2006.

[134] S. Tasharofi, F. Raja, F. Oroumchian, and M. Rahgozar, "Evaluation of statistical part of speech tagging of persian text," in *2007 9th International Symposium on Signal Processing and Its Applications*, pp. 1–4, IEEE, 2007.

[135] W. Anwar, X. Wang, L. Li, and X.-L. Wang, "A statistical based part of speech tagger for urdu language," in *2007 international conference on machine learning and cybernetics*, vol. 6, pp. 3418–3424, IEEE, 2007.

[136] A. Kadim and A. Lazrek, "Parallel hmm-based approach for arabic part of speech tagging.," *Int. Arab J. Inf. Technol.*, vol. 15, no. 2, pp. 341–351, 2018.

[137] D. E. Cahyani and M. J. Vindiyanto, "Indonesian part of speech tagging using hidden markov model–ngram & viterbi," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 353–358, IEEE, 2019.

[138] Z. Z. Linn and P. B. Patil, "Part of speech tagging for kayah language using hidden markov model," in *2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)*, pp. 228–233, IEEE, 2019.

[139] J. Singh, N. Joshi, and I. Mathur, "Development of marathi part of speech tagger using statistical approach," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1554–1559, IEEE, 2013.

[140] S. Mohammed, "Using machine learning to build pos tagger for under-resourced language: the case of somali," *International Journal of Information Technology*, vol. 12, no. 3, pp. 717–729, 2020.

[141] L. Yimin and H. De-gen, "Chinese part-of-speech tagging based on full second-order hidden markov model," *Computer Engineering*, vol. 31, no. 10, pp. 177–179, 2005.

[142] S. K. Daimary, V. Goyal, M. Barbora, and U. Singh, "Development of part of speech tagger for assamese using hmm," *International Journal of Synthetic Emotions (IJSE)*, vol. 9, no. 1, pp. 23–32, 2018.

[143] S. Bandyopadhyay and A. Ekbal, "Hmm based pos tagger and rule-based chunker for bengali," in *Advances in pattern recognition*, pp. 384–390, World Scientific, 2007.

[144] M. Banko and R. C. Moore, "Part-of-speech tagging in context," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 556–561, 2004.

[145] M. H. Davis, *Markov models & optimization*. Routledge, 2018.

[146] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[147] H. Xiang and Z. Ou, "Crf-based single-stage acoustic modeling with ctc topology," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5676–5680, IEEE, 2019.

[148] R. Ghaffari, M. Golpardaz, M. S. Helfroush, and H. Danyali, "A fast, weighted crf algorithm based on a two-step superpixel generation for sar image segmentation," *International Journal of Remote Sensing*, vol. 41, no. 9, pp. 3535–3557, 2020.

[149] S. Girisha, M. M. Pai, U. Verma, and R. M. Pai, "Semantic segmentation with enhanced temporal smoothness using crf in aerial videos," in *2021 IEEE Madras Section Conference (MASCON)*, pp. 1–5, IEEE, 2021.

[150] A. Ekbal, R. Haque, and S. Bandyopadhyay, "Bengali part of speech tagging using conditional random field," in *Proceedings of seventh international symposium on natural language processing (SNLP2007)*, pp. 131–136, Citeseer, 2007.

[151] S. L. Pandian and T. Geetha, "Crf models for tamil part of speech tagging and chunking," in *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy: 22nd International Conference, ICCPOL 2009, Hong Kong, March 26-27, 2009. Proceedings 22*, pp. 11–22, Springer, 2009.

[152] K. Nongmeikapam, L. Nonglenjaoba, Y. Nirmal, S. Bandhyopadhyay, *et al.*, "Improvement of crf based manipuri pos tagger by using reduplicated mwe (rmwe)," *arXiv preprint arXiv:1111.2399*, 2011.

[153] T. D. Singh, A. Ekbal, and S. Bandyopadhyay, "Manipuri pos tagging using crf and svm: A language independent approach," in *proceeding of 6th International conference on Natural Language Processing (ICON-2008)*, pp. 240–245, 2008.

[154] A. K. Barman, J. Sarmah, and S. K. Sarma, "Pos tagging of assamese language and performance analysis of crf++ and fntbl approaches," in *2013 UKSim 15th International Conference on Computer Modelling and Simulation*, pp. 476–479, IEEE, 2013.

[155] A. K. Ojha, P. Behera, S. Singh, and G. N. Jha, "Training & evaluation of pos taggers in indo-aryan languages: a case of hindi, odia and bhojpuri," in *the proceedings of 7th language & technology conference: human language technologies as a challenge for computer science and linguistics*, pp. 524–529, 2015.

[156] S. Ghosh, S. Ghosh, and D. Das, "Part-of-speech tagging of code-mixed social media text," in *Proceedings of the second workshop on computational approaches to code switching*, pp. 90–97, 2016.

[157] N. Suraksha, K. Reshma, and K. S. Kumar, "Part-of-speech tagging and parsing of kannada text using conditional random fields (crfs)," in *2017 International Conference on Intelligent Computing and Control (I2C2)*, pp. 1–5, IEEE, 2017.

[158] Z. Nasim, S. Abidi, and S. Haider, "Modeling pos tagging for the urdu language," in *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, pp. 1–6, IEEE, 2020.

[159] E. Minkov, R. C. Wang, and W. Cohen, "Extracting personal names from email: Applying named entity recognition to informal text," in *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pp. 443–450, 2005.

[160] T.-V. T. Nguyen and A. Moschitti, "Structural reranking models for named entity recognition," *Intelligenza Artificiale*, vol. 6, no. 2, pp. 177–190, 2012.

[161] P. Thangzikpuia, "Mizo tawng grammar(based on its usage and unique features)," 2019.

[162] K. Lalzarzova, "Mizo tawng grammar & composition," *R.K Lalhluna, Lunglei, Mizoram*, 2016.

[163] R. Lalhluna, "Cinque foils – zo Ṭawng grammar," *K. Sangzawna, Aizawl, Mizoram*, 2014.

[164] M. B. of School Education, "Mizo awng ziah dan," 2020.

[165] D. Modi, N. Nain, M. Nehra, *et al.*, "Part-of-speech tagging for hindi corpus in poor resource scenario," *Journal of Multimedia Information System*, vol. 5, no. 3, pp. 147–154, 2018.

[166] R. Narayan, S. Chakraverty, and V. Singh, "Neural network based parts of speech tagger for hindi," *IFAC Proceedings Volumes*, vol. 47, no. 1, pp. 519–524, 2014.

[167] F. Jahara, A. Barua, M. Iqbal, A. Das, O. Sharif, M. Hoque, and I. Sarker, "Towards pos tagging methods for bengali language: a comparative analysis," in *International Conference on Intelligent Computing & Optimization*, pp. 1111–1123, Springer, 2020.

[168] S. Sharma and G. Lehal, "Using hidden markov model to improve the accuracy of punjabi pos tagger," in *2011 IEEE International Conference on Computer Science and Automation Engineering*, vol. 2, pp. 697–701, IEEE, 2011.

[169] C. Jobanputra, N. Parikh, V. Vora, and S. K. Bharti, "Parts-of-speech tagger for gujarati language using long-short-term-memory," in *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, pp. 1–5, IEEE, 2021.

[170] C. Tailor and B. Patel, "Hybrid pos tagger for gujarati text," in *International Conference on Soft Computing and its Engineering Applications*, pp. 134–144, Springer, 2020.

[171] V. Gamit, R. Joshi, and E. Patel, "A review on part-of-speech tagging on gujarati language," *International Research Journal of Engineering and Technology (IRJET)*, 2019.

[172] A. Priyadarshi and S. K. Saha, "Towards the first maithili part of speech tagger: Resource creation and system development," *Computer Speech & Language*, vol. 62, p. 101054, 2020.

[173] S. Daimary, V. Goyal, M. Barbora, and U. Singh, "Development of part of speech tagger for assamese using hmm," *International Journal of Synthetic Emotions (IJSE)*, vol. 9, no. 1, pp. 23–32, 2018.

[174] D. Pathak, S. Nandi, and P. Sarmah, "Aspos: Assamese part of speech tagger using deep learning approach," in *2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1–8, IEEE, 2022.

[175] M. Tham, "A hybrid pos tagger for khasi, an under resourced language," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, 2020.

[176] S. Warjri, P. Pakray, S. Lyngdoh, and A. Maji, "Part-of-speech (pos) tagging using deep learning-based approaches on the designed khasi pos corpus," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 3, pp. 1–24, 2021.

[177] P. Vaishali, K. Kalpana, and N. MC, "A rule-based approach for marathi part-of-speech tagging," in *ICT with Intelligent Applications*, pp. 773–785, Springer, 2022.

[178] P. Antony and K. Soman, "Parts of speech tagging for indian languages: a literature survey," *International Journal of Computer Applications*, vol. 34, no. 8, pp. 0975–8887, 2011.

[179] B. Harish and R. Rangan, "A comprehensive survey on indian regional language processing," *SN Applied Sciences*, vol. 2, no. 7, pp. 1–16, 2020.

[180] D. Kumar and G. Josan, "Part of speech taggers for morphologically rich indian languages: a survey," *International Journal of Computer Applications*, vol. 6, no. 5, pp. 32–41, 2010.

[181] "The linguistic data consortium for indian languages (ldc-il)." `https://www.ldcil.org/default.aspx`.

[182] K. Lalzarzova, *Mizo Tawng Grammar Composition*. R.Lalrawna, 2016.

[183] P. Thangzikpuia, *Mizo Tawng Grammar(Based on its usage and unique features)*. P.C. Thangzikpuia, 2019.

BIO-DATA OF THE CANDIDATE

Name of Candidate	: Morrel V.L. Nunsanga
Date of Birth	: 19/01/1980
Phone	: 9862596269
email	: morrelhmar@mzu.edu.in
Permanent Address	: S/o L.H. Chuailova Dinthar, Aizawl, Mizoram Pin: 796001
Married	: Yes
Educational Details	
(a) M.Tech	: Tezpur University
(b) Ph. Course Work	: Mizoram University
Present Occupation	: Assistant Professor, : Department of Information Technology
Organization	: Mizoram University

LIST OF PAPERS/PATENT BASED ON THESIS

Journals:

1. Nunsanga, Morrel VL, Partha Pakray, C. Lallawmsanga, and L. Lolit Kumar Singh. "Part-of-Speech Tagging for Mizo Language Using Conditional Random Field." *Computación y Sistemas,* 25, no. 4 pp 803-812 , 2021. ISSN: 1405-5546 (ESCI/Scopus)

2. Nunsanga, M. V., Chhawnchhek, L., Pakray, P., L. Lolit Kumar Singh. "An Emperical Study on POS tagging for Mizo Language." *Science and Technology Journal,* Vol. 10. Issue 2 July, 2022, ISSN: 2321-3388 (UGC Care list)

3. Lawmsanga, Nunsanga, Morrel VL, Partha Pakray, and L. Lolit Kumar Singh. "POS tagging for Mizo language: Unique features and challenges." *Mizo Studies,* Vol X. No. 1, 76-88 (January - March 2021) ISSN: 2319 6041 (UGC Care list)

4. Nunsanga, Morrel VL, Partha Pakray, T. Sonalika Devi. "Enhancing HMM-based POS tagger for Mizo language", in *Journal of Intelligent & Fuzzy Systems*. Vol. 45, no. 6, pp. 11725-11736, December 2023, ISSN: 1875-8967 (E), 1064-1246 (P) (SCIE, IF 2023 : 2.0)

Conference:

1. Nunsanga, Morrel VL, Partha Pakray, Mika Lalngaihtuaha, and L. Lolit Kumar Singh. "Stochastic based part of speech tagging in Mizo language: Unigram and Bigram hidden Markov model." In *Edge Analytics: Select Proceedings of 26th International Conference—ADCOM 2020,* pp. 711-722. Singapore: Springer Singapore, 2022. (Scopus)

2. Nunsanga, Morrel VL, Partha Pakray, Mika Lalngaihtuaha, and L. Lolit Kumar Singh. "Part-of-speech tagging in Mizo language: A preliminary study." In *Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2020,* pp. 625-635. Springer Singapore, 2021.

Patent:

1. Nunsanga, M. V., Chhawnchhek, L., Pakray, P., L. Lolit Kumar Singh. "A System For Studying And Performing The Part-of-Speech Tagging For Mizo Language" - *German Utility Patent,* 18 July 2022, No.: 202022104028

PARTICULARS OF THE CANDIDATE

NAME OF CANDIDATE	: MORREL V.L. NUNSANGA
DEGREE	: PH.D
DEPARTMENT	: INFORMATION TECHNOLOGY
TITLE OF THE THESIS	: ANALYSIS OF PART OF SPEECH TAGGING FOR MIZO LANGUAGE
DATE OF ADMISSION	: 03.07.2018
APPROVAL OF RESEARCH PROPOSAL	
1. DRC	: 4th April 2019
2. BOS	: 23rd April 2019
3. SCHOOL BOARD	: 30th May 2019
MZU REGISTRATION NO.	: 1807566
PH.D REGISTRATION NO.	: MZU/Ph.D/1298 of 03.07.2018
EXTENSION	: NA

(DR. R. CHAWNGSANGPUII)

Head

Department of Information Technology

ABSTRACT

ANALYSIS OF PART OF SPEECH TAGGING FOR MIZO LANGUAGE

AN ABSTRACT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

MORREL V.L. NUNGSANGA

MZU REGN NO : 1807566

Ph.D. REGN NO : MZU/Ph.D/1298 OF 03.07.2018

DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF ENGINEERING AND TECHNOLOGY

SEPTEMBER, 2023

ANALYSIS OF PART OF SPEECH TAGGING FOR MIZO LANGUAGE

BY

MORREL V.L. NUNGSANGA
DEPARTMENT OF INFORMATION TECHNOLOGY

Supervisor : Prof. L. Lolit Kumar Singh
Joint-Supervisor : Dr. Partha Pakray

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Information Technology of Mizoram University, Aizawl

ABSTRACT

In the realm of natural language processing, accurate part-of-speech tagging serves as a fundamental building block, facilitating various language understanding tasks. This thesis delves into the intricate domain of part-of-speech tagging for the Mizo language, with a focus on designing and implementing a high-accuracy tagger. The primary aim of this research is to establish a reliable computational framework capable of proficiently assigning grammatical tags to words in Mizo sentences. The study encompasses a multifaceted approach, combining computational analysis, exploration of methodologies, resource development, grammar-based insights, and novel tagger design.

The research unfolds through a multifaceted approach encompassing diverse stages. It commences with an in-depth computational analysis that sheds light on the distinct characteristics of the Mizo language. To comprehend the complex nuances of the Mizo language, an extensive computational analysis was conducted, unraveling its unique linguistic traits, contextual cues, and distinctive patterns. This analysis led to the identification of essential linguistic key features, which lay the foundation for an effective part-of-speech tagging system.

The research involved an exploration of diverse methodologies and techniques prevalent in the realm of part-of-speech tagging. Traditional rule-based approaches, statistical models, and hybrid techniques are assessed in the context of Mizo. This exploration facilitated the identification of methodologies best suited to the intricate characteristics of the Mizo language. The study also considers the challenges posed by the limited availability of labeled data for Mizo and explores techniques for domain adaptation and data augmentation to alleviate this limitation.

A pivotal contribution is the creation of a robust linguistic resource tailored to Mizo. The establishment of a comprehensive tagset, encompassing the diverse grammatical categories and linguistic phenomena of the language, lays the foundation for accurate tagging. In tandem, an extensively annotated corpus of Mizo text is curated, representing a wide array of domains and contexts. This dataset not only serves as the bedrock for training and validating the part-of-speech tagger but also provides a valuable asset for advancing broader language processing applications for Mizo.

An innovative facet of this work is the development of a precise regular expression framework that transcends mere pattern matching. This framework, informed by linguistic insights gained from the computational analysis, captures intricate syntactic patterns and contextual cues within Mizo sentences. By encapsulating these linguistic nuances, the regular expressions enrich the tagging process, providing the tagger with additional linguistic context and enhancing its accuracy.

In sum, this research makes substantial strides in the realm of Mizo language processing. Through a meticulous examination of its linguistic intricacies, grammatical structures, and contextual nuances, the study opens pathways for the development of cutting-edge part-of-speech taggers fine-tuned to the idiosyncrasies of the Mizo language. As a result, this work not only contributes to the enrichment of language technology but also deepens our comprehension of this unique language's underlying mechanics. The implications of this research extend beyond part-of-speech tagging, laying the groundwork for enhanced language understanding and downstream applications for the Mizo language.