2020-06-22

PART OF SPEECH TAGGING FOR AWNGI LANGUAGE

BIRHANE, WONDMANEH GETAHUN

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

SCHOOL OF RESEARCH AND POSTGRADUATE STUDIES

COMPUTING FACULTY

DEPARTMENT OF INFORMATION TECHNOLOGY

PART OF SPEECH TAGGING FOR AWNGI LANGUAGE

MSc. Thesis

BY

BIRHANE WONDMANEH GETAHUN

BAHIR DAR, ETHIOPIA

June 22, 2020

Part of Speech Tagging for Awngi Language

By

Birhane Wondmaneh Getahun

A thesis submitted to the school of Research and Graduate Studies of Bahir Dar Institute of Technology, in partial fulfillment of the requirements for the degree of Master of Science (MSc) in Information Technology

Advisor Name: Assefa M. (Assistant Professor)

Bahir Dar, Ethiopia

June 22, 2020

DECLARATION

I, the undersigned, declare that the thesis comprises my own work. In compliance with internationally accepted practices, I have acknowledged and refereed all materials used in this work. I understand that non-adherence to the principles of academic honesty and integrity, misrepresentation/ fabrication of any idea/data/fact/source will constitute sufficient ground for disciplinary action by the University and can also evoke penal action from the sources which have not been properly cited or acknowledged.

Name of the student Birhane Wondmeneh Signature _____

Date of submission: 22/06/2020

Place: Bahir Dar University

This thesis has been submitted for examination with my approval as a university advisor.

Advisor Name: Assefa Misganaw

Advisor's Signature: _____

Bahir Dar University

Bahir Dar Institute of Technology-

School of Research and Graduate Studies

Faculty of computing

THESIS APPROVAL SHEET

Student:

Birhane Wondmaneh _____ _____ 22/06/2020

Name Signature Date

The following graduate faculty members certify that this student has successfully presented the necessary written final thesis and oral presentation for partial fulfillment of the thesis requirements for the Degree of Master of Science in Information Technology.

Approved By:

Advisor:

Assefa Misganaw(Assi prof) _____ 22/06/2020

Name Signature Date

External Examiner:

Yaregal Assabie (PhD) _____ 18/06/2020

Name Signature Date

Internal Examiner:

Mekuanint Agegnehu (PhD) _____ 22/06/2020

Name Signature Date

Chair Holder:

Abrham Debasu(Assi prof) _____ 22/06/2020

Name Signature Date

Faculty Dean:

Belete Biazen _____ 22/06/2020

Name Signature Date

ACKNOWLEDGEMENT

First and for most I would like to thank almighty God for giving me the strength and perseverance to progress one step forward in doing this research. Next I thank Mr Assefa M.(Assi Professor) for his fruit full advises and suggestions; particularly by giving a general clue on how to do a scientific research. Last but not least my gratitude directly goes to the Awngi language experts Mr. Ayenew W. and Mr. Tesfaye for helping me day and night in order to preparing and tagging the data set. Finally, I would like to thank my family and friends for motivating me by giving moral for completing this research work.

ABSTRACT

Natural Language Processing (NLP) has emerged as a means of increasing computers capability to understand natural languages, by which most of human knowledge is recorded. Part-of- Speech (POS) tagging is one of the tasks of NLP, which is used for labeling or classification on every word of a text with its correct part of speech category like noun, verb, adjective, adverb, preposition, conjunction etc. based on its definition and context of adjacent and related word. Awngi language is categorized under a Cushitic language family which is spoken by more than 1.5 million people in Amhara and some Parts of Benishangul Gumuz Regional states. This language has been one of the under-resourced languages both in terms of electronic resources and processing tools. In this regard, different natural language processing tasks are left for researchers for investigation. Among the different research areas of NLP, we used to focus on part of speech tagging since it is the primary and fundamental work. The output of POS tagging will be used as an input for grammar checker, spell chucker, information extraction, information retrieval, speech synthesis, parsing of text, semantic processing and e.t.c.

The main motivation for this resource is to obtain data for training automatic taggers with machine learning approach. Hence, we take machine learning considerations into account during tagset design and present training experiments as part of this paper.

Awngi language corpus is not available in organized manner. As a result, with the help of experts, we have prepared and tag the training data set manually. The data was collected from Injibara elementary and high school Awngi text books, from Amhara Mass Media Agency Awngi radio and Television program, as well as from Awi zone administration office. In order of reducing the complexity of part of speech tagging we have used the Hidden Markov model statistical approach. For the selected approach N –gram Viterbi algorithm is used for tagging purpose.

For result simulation, we have used python programming language with a tenfold cross validation evaluation mechanism and finally the average accuracy of the tagger becomes 91.3% which is significant for the future researchers who want to investigate NLP researches on Awngi language in particular and on other local languages in general.

TABLE OF CONTENTS

CONTENTS PAGES

LIST OF ABBREVIATIONS

POST Part of Speech Tagging

DBU Bahir Dar University

HMM Hidden Markov Model

ANN Artificial Neural Network

FCV Fold Cross Validation

NLP Natural Language Processing

IR Information Retrieval

IE Information Extraction

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

1.1. Background of the study

Natural Language Processing (NLP) is a branch of computational linguistics which is concerned with automated, computer processing of natural language such as speech acts or texts. It concerns to process and understand natural language using computers and is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. Thus, it performs useful tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech (D. Jurafky ,J. H. Martin, 2006). It includes techniques like word stemming (removing suffixes) or a related technique, lemmatization (replacing an inflected word with its base form), multiword phrase grouping, synonym normalization, part-of-speech (POS) tagging (elaborations on noun, verb, preposition etc.), word-sense disambiguation, and role determination (e.g. subject and object) (Anne Kao,Steve Poteet:). In addition it has many applications including machine translation, speech recognition, question answering, information retrieval system and parts of speech tagging. The above definition uses for all natural languages to develop the part of speech tagging.

These days, due to growth of scientific and technical advances, there is large amount of information that is retained and processed by business and organizations. Some of such information is stored in the form of text. Processing and retrieving useful information from such hugely available data and information is very difficult for human beings. To tackle such problem professionals and scientists from different area of studies like Artificial intelligence, information retrieval, natural language processing, data mining etc. started to conduct research that mainly focuses on helping computers understand natural languages (James, 1995).

Therefore, as a solution to problems of helping computers understand natural languages, Natural Language Processing (NLP) has emerged as a means of increasing computers capability to understand natural languages, by which most of human knowledge is

recorded. It is used to design and implement tools, techniques and frameworks that enable computers communicate effectively with each other and with humans. As scientific study encompasses a set of related disciples like psycholinguistic, linguistic and computational linguistic and other related fields to study and design effective components like morphological analyzer, syntax parser, semantic analyzer, speech recognizer, part of speech (POS) tagger and many more application that can help computers understand text, sounds, images and other forms of information just like that of human being (Daniel J. et al, 2018). NLP applications take advantage of machine learning strategies in order to analyze large amounts of textual data. One of such tasks to be performed on textual data is Part-of- Speech (POS) tagging, which is used for labeling or classification of every word of a text with its correct part of speech category like noun, verb, adjective, adverb, preposition, interjection etc. based on its definition and context of adjacent and related word (Marcos G. et al , 2014).

So far many POS tagging researches have been done and different approaches have been used for POS tagging, where the well-known ones are rule-based, stochastic, Artificial Neural Networks and Hybrid Approach. Rule-based taggers, as their name implies, (Brill, Eric, 1992) strive to assign a tag to each word using a set of handwritten rules that might be specified by language experts or machine learned rules. These rules could determine, for example, that a word following a determiner and an adjective must be a noun. In this approach, what researchers should do is set and check rules properly with the help of language professionals or to make the taggers learn rule during the training phase of the system as it is in the case of the Brill tagger (Brill, Eric, 1992). The stochastic (statistic or probabilistic) approach (Getachew M. , 2001)uses a training corpus to pick the most probable tag sequence for a word sequence in a given sentence to be tagged. Some stochastic methods are based on first order or second order Markov models and some are based on a few other techniques which use probabilistic approach for POS Tagging, such as the Tree Tagger (Schmid H. , 1994). Artificial Neural Network (ANN) uses a training corpus and adaptively learns properties of words to pick the appropriate tag for a word in a given sentence (Solomon, 2008).

Finally, the hybrid approach may either combine the rule-based approach and statistical approach or the rule based approach and the Artificial Neural Network. The hybrid of

rule based and stochastic approach for example may pick the most likely tag based on a training corpus and then applies a certain set of rules to see whether the tag should be changed to another tag or not. Besides it saves any new rules that it has learnt in the process, for future use. One example of an effective tagger in this category is the Brill Tagger (Brill, Eric, 1992).

1.2. Statement of the problem

Ethiopia is the home of more than 85 nations, nationalities and peoples speaking different languages. The nature of those languages may vary in terms of structure like morphology, phonology, syntactical and lexical rules. Among the different languages available in Ethiopia, Awngi is a Central Cushitic language spoken by at least 1.5 million people in an extensive area in northwest Ethiopia, including all of Awi Zone, but also some areas of the Metekel Zone of the Benishangul-Gumuz National Regional State, and various places in the Alefa and K'wara Woredas of the North-Gonder Zone of the Amhara National Regional State. The Alefa and K'wara varieties have sometimes been called Kunfal, but are dialects of the Awngi language. (Tsegaye, 2013)

Researchers tried to investigate in the area of natural language processing specifically on POS tagging for Ethiopian languages like Amharic, Affan Oromo, wolaita and Tigregna etc. Now a day, using this language, there is Educational curriculum starting from grade primary to secondary schools, it is the working language for Awi zone and a lot of daily radio and television news are broadcasting from the Amhara Regional state mass media agency.

The daily increasing need of the society will lead to use technological practices like that of natural language processing. From the different types of NLP tasks the absence of POS Tagger system limits researches concerning the NLP of Awngi language such as parsing (syntactic and semantic), machine translation, sentence grammar checker, spell checker, speech synthesis etc as it is used as a pre-processing component for the aforementioned NLP applications.

Awngi language uses geez writing system. The language uses both kinds of morphologies, i.e. inflectional and derivational. As pointed out by (Willet, 2002) depending on the morphological complexity of a language, both inflectional and derivational morphologies can result in very large numbers of variants of part of speech tagging for a single word. As a result, word form variations can have a strong impact on the effectiveness of Part of speech tagging systems and on morphological analysis tools. In Awngi, inflectional and derivational conditions, the morphological structure involves suffixing.

Consider the following Awngi words example (A, B)

A. አውኺስስታ አምሀርኺስ ዙሚትሻስ

English Meaning Let's talk by Awgi and by Amharic

Amharic meaning በአዊይኛ እና በ አማርኛ እንነጋገር

Unlike Amharic and English language Awngi morphology has no prefix rather it will consolidate after suffix or after root word.

When we see the morphology of the above single word አውኺስስታ

Root word = አውኺ Suffix = ስ Prefix added after suffix = ስታ

አምሀርኺስ Root word= አምሀርኺ

Prefix found at the end of the word is the letter ስ

B. ኬቴሙዝኩዋንትካ

Root word ኬቴሙ

Root word ዝኩዋን

Suffix ት

Prefix added after suffix ካ

The Awngi language and grammatical structure words are created by adding suffixes to a basic stem. It uses an extensive concatenation of suffixes. This characteristic of the language make a very short stem into a long word. As it is evidenced by different authors, such nature of the language indicates the complexity of the language's morphology. Hence, conducting a research on Part of speech tagging worth paramount significant.

Finally the researcher is supposed to answer the following research questions.

1.2.1. Research Questions

- How can we investigate the basic structures and rules of Awngi language for easily tagging of part of speech?
- How to identify the Awngi tag sets that are used for labeling Awngi words?
- How to test and analyze the Awngi POS tagger system performance?

1.3. Objectives of the study

1.3.1. General objective

The general objective of this research work is to develop Part of Speech Tagger for Awngi language.

1.3.2. Specific objectives

In order to achieve the above general objective, the proposed research will accomplish the following specific objectives:

- Review, Analyze and study the structure of the Awngi sentence.
- Study the morphological property of the language to identify properties useful to POS Tagger.
- Collect and design corpus for training and testing of the system
- Design and model a POS Tagger for the language.
- Develop Awngi POS Tagger Prototype.
- Test the system performance

1.4. Methodology of the study

Here are the possible methodologies which are used to investigate part of speech tagging for awngi language.

1.4.1. Literature review

For understanding the exact research gaps and for using the appropriate research methods, tools and techniques for part of speech tagging, literatures are used as a methodology.

1.4.2. Data collection

The primary as well as secondary data has been collected from Injibara elementary and high school Awngi text books, from Amhara Mass Media Agency Awngi Radio and Television program, and from Awi zone administration office. The content of the data was both in the form of softcopy and in hardcopy. In order to understand the language structure, to determine the tag sets and to tag the corpus manually, continuous discussion with linguists and experts in the area of Awngi language was the basic task for the researcher.

1.4.3. Modeling

Based on the amount of dataset and the linguistic experts we have available budget and time, the researcher adopted HMM. The HMM based Tagger relies on the statistical property of words along with part of speech categories. Such statistical property can be distributional probability of words with tags which can be obtained during the training phase of the system.

1.4.4. Tools and Implementation

To conduct research on POS Tagger for Awngi text, an open source Natural Language ToolKit (NLTK) and Python programming language will be used. The rationale behind the choice of these two tools is that they are suitable for processing different NLP tasks. NLTK is an open source tool that contains open source python modules, linguistic data and documentation for research and development in natural language processing (python.org, 2018) Python is an easy to learn but powerful programming language especially for text processing in NLP applications. It has efficient high level data structures and a simple but effective approach to object-oriented programming (python.org, 2018).

1.5. Scope and limitations of the study

1.5.1. Scope of the study

As it has been discussed on the above portions of the research in Awngi language there is no available dataset that can be used as an input for natural language processing research investigations. As a result, the scope of this research is to prepare necessary data set for testing and training purpose to investigate part of speech tagging for Awngi text using small manually tagged corpus by using Hidden Markov model statistical approach. At the end of the research work, the performance of the tagger has been evaluated using tenfold cross validation which is the most appropriate evaluation criteria's for small dataset.

1.5.2. Limitations of the study

Awngi language orthography was prepared since 1990's. As a result, it is under resourced language and was difficult to prepare the training data set. In addition Lack of Awngi language experts to tag large amount of data set for training set was a challenge, another problem was lack of literatures and related works on Awngi language for example, some of Awngi root word is not determined clearly so that the researcher faced a difficulty for tagging.

1.6. Significance of the study

As POS tagging is a basic and the first NLP task for any language, it can be used as an input for other high level NLP tasks such as sentence grammar checker, parsing, machine translation, word sense disambiguation etc. This research can also be used as a reference for other researchers who are interested to work on Awngi language.

CHAPTER TWO: LITERATURE REVIEW AND RELATED WORK

2.1. Literature review

So far a lot of POS tagging researches have been done and there are different approaches used. Among them the well-known ones are rule-based, stochastic, Artificial Neural Networks and Hybrid Approach. Rule-based taggers, as their name implies, (Bird, 2006)target to assign a tag to individual word using a set of handwritten rules that might be specified by language experts or machine learned rules. These rules determine that a word following a determiner and an adjective must be a noun. In this approach, what researchers should do is set and check rules properly with the help of language professionals or to make the taggers learn rule during the time of training phase of the system as it is in the case of the Brill tagger (Brill, 1995).

The other one is the stochastic (statistical or probabilistic) approach (Fahim Muhammad et al., 2006) uses a specified training corpus to pick the most probable tag sequences for a word sequence in a given sentence to be tagged. Some stochastic methods are based on first order or second order Markov models and some of them are based on a few other techniques which use probabilistic approach in order to POS Tagging, such as the Tree Tagger (Schmid, Helmut, 1994).

The other one is Artificial Neural Network (ANN) which uses a training corpus and adaptively learns properties of words to pick the appropriate tag for a token in a given sentence (Solomon A. , 2008)e.

Finally, the hybrid approach may either combine the statistical approach and rule-based approach or the Artificial Neural Network and the rule based approach. The hybrid of stochastic approach and rule based for instance may pick the most likely tag based on a training corpus and then applies a certain set of rules to see whether the tag should be changed to another tag or not. Besides it saves any new rules that it has learnt in the process, for future use. Among them the One example of an effective tagger is the Brill Tagger (Hussen, 2010).

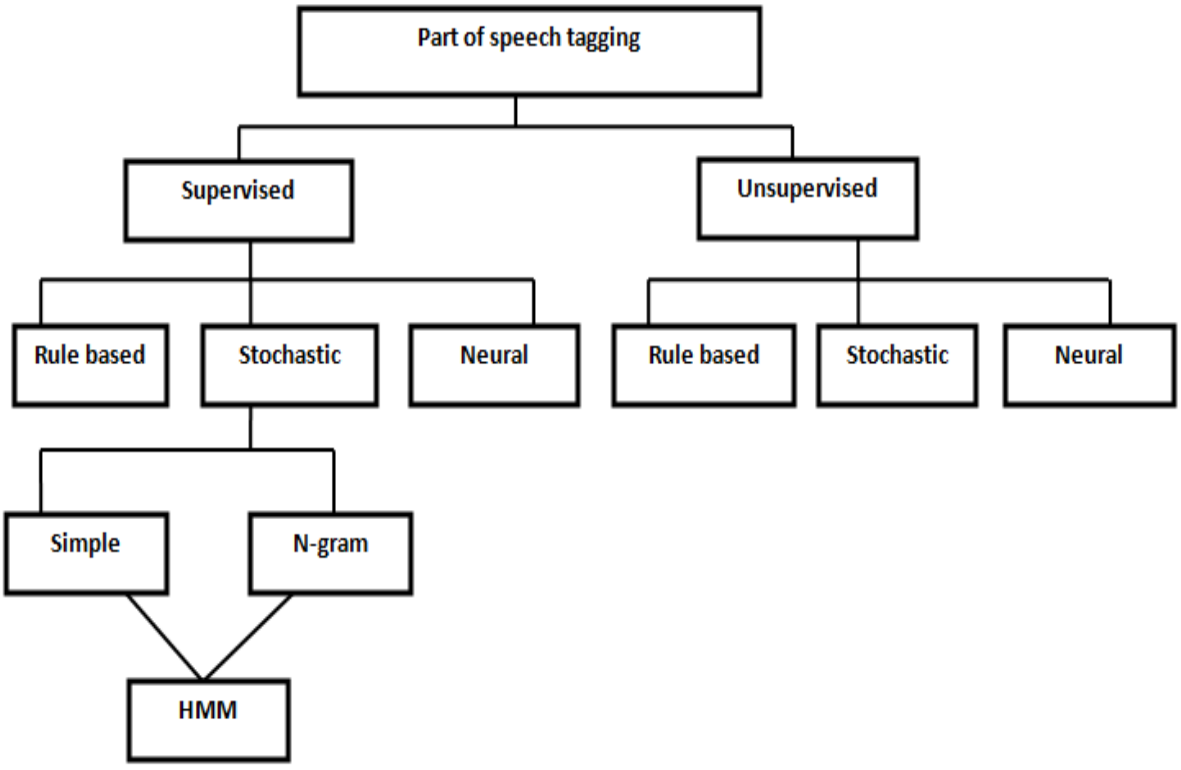


Figure 2.1- The Common Methods for the POS taggers **(HASAN MUAIDI , Levenberg-Marquardt, 2014)**

2.1.1. Approaches of Part of Speech tagging

2.1.1.1. Rule Based Approach

The rule based approach uses predefined rules to in order to disambiguate tags of words. The rules are based on knowledge of the specific language experts which may consist of a large number of morphological, lexical and syntactical information. These rules can be obtained manually that are handcrafted by linguistic professionals or through machine learning. The former way of getting rules is tedious, time taking since it requires linguistic professionals to manually set rules. furthermore, it is inconsistent and subjective as it is determined by the understanding of one or more linguistic specialists and their skill and knowledge of the specific language (Solomon, 2008), the later way of obtaining rules as it has been explained , transformation-error driven approach, is from a training corpus. That means a model is made to automatically learn and store rules (also called Brill transformations) from the training corpus to be provided. There is no way of

specifying the rules manually by linguistic professionals, what is needed is the tagged corpus as an input to automatically drive its own rules so called transformations in Brill tagger (Brill, 1995).

In addition to what has been discussed above, the rule based approach uses contextual information to assign tags to unknown words. These rules are often known as context frame rules. A context frame rule can be, if an unknown word is preceded by determiner and followed by a noun, its correct tag is adjective. Moreover morphological information can be used as a rule to aid in the tagging process. One specific example in this case is, if a token ends with „-ing" and is preceded by a verb the most probable tag of the token is verb. (Levent Altunyurt , Zihni Orhan, 2006)

In fact, adding a certain rule to a system may involve over-generation, i.e., one extra rule can result in more harm to the accuracy of general tagging in machine-learning rule-based approach. The manual rule-based tagging system has also the aforementioned limitations. Therefore, to conclude, the rule-based approaches are time-consuming and require a great knowledge of a specific language experts being tagged. But it is also possible to find some advantages of the rule based approach that is listed in the work of (Brill, Eric, 1992). These are: Robustness, a vast reduction in stored information required the lucidity of a small set of meaningful rules, ease of finding and implementing improvements to the tagger, and also better portability from one tag set or corpus type to another. Rule based approach has its own advantages. Among them it requires only small amount of training data, Can be used with both well-formed and ill-formed input, High quality based on solid linguistic and it is useful for limited domain. In addition it has also several disadvantages namely, it relies on hand-constructed rules that are to be acquired from language specialists and construction of these rules is tedious and time consuming, development could be very time consuming, some changes may be hard to accommodate and not easy to obtain high coverage of the linguistic knowledge.

2.1.1.2. Stochastic Approach

The stochastic approach also called statistical approach is based on a probabilistic pattern in order to assign a probable part of speech tag to a given text from a given training text corpus. The goal of any stochastic approach is to pick the most probable tag for a word

from its context and its neighbors (Bird, 2006). They can build a probability matrix that stores the probability of an individual word belonging to a certain part of speech and its distributional probability. They can use this distributional probability to tag words that are in the input sentence but not in the training corpus.

The probabilities are estimated from a tagged training corpus or an untagged corpus. Stochastic tagging techniques can be of two types depending on the training data. Supervised Statistical tagging techniques use tagged corpus for their training though it requires large amount of tagged data so that high level of accuracy can be achieved. Unsupervised Statistical techniques, on the other hand, are those which do not require a pre-tagged corpus but instead use sophisticated computational methods to automatically induce word groupings (i.e. tag sets), and based on these automatic groupings, they calculate the probabilistic values needed by statistical taggers.

The basic idea of this approach is to find the probability (p) of a word along its tag from a given sentences or text which can be represented mathematically as:
P(wi,Ti)|<S>Where Wi, Ti are the i^{th} word and the i^{th} tag in the input sentence or text <S>.
The stochastic approach may use the most frequent tag, N-gram analysis or Hidden Markov Models to disambiguate tag of words which can be derived from the above mathematical representation.
The most frequent tag model as the name implies tries to pick the most frequent tag for a given word in a given sentence. This model is the simplest model in the stochastic approach as it simply finds the most frequent tag from the training corpus. This can be done by counting the occurrence of the specific word Wi associated with a tag Ti and dividing it by the total occurrence of the word Wi in the training corpus which can be represented mathematically as:

$$P(Ti|Wi) = \frac{\text{count of (Ti,Wi)}}{\text{count of Wi}}$$

This implies that the most frequent tag model computes the probabilities of observing each word attached with every part of speech tags during the training phase. Then in the tagging process of new text, it will pick the tag with most probable tag for that word. This

model has very clear limitations as it does not try to look into the sentence structure. This can be solved using the N-gram model that deals with local context of words in a sentence.

The N-gram model is a mechanism of dealing local context of words in a given sentence. It is originally conceived as a technique of predicting the next element of a sequence given only the N-1 previous elements (D. Jurafky ,J. H. Martin, 2006). The elements can be sequence of words, tags or both words and their corresponding tags. This implies that it can be used for finding the tag sequence probabilities (probability of a tag Ti given the previous N-1 tags (P(Ti|Ti-1,Ti-2, …, Tn)) or word sequence probabilities ((P(Wi|Wi-1,Wi-2, …, Wn)). Moreover, the n-gram can also be used for finding the probability of the tag of a current word given the previous n words (P(Ti,Wi |Wi-1, Wi-2, …, Wn)). It solves the problem of the most frequent tag model which ignores the local context of words. This model decides the appropriate tag for a word by computing the probability that it occurs within the n-previous tags, where the value of N is considered to be 1, 2, or 3 for practical purposes (Mamo, 2009). These are known as unigram, bigram and trigram models respectively.

The Hidden Markov Model (HMM) is the most widely used model for part of speech tagging under the stochastic approach (Sandipan Dandapat, Sudeshna Sarkar ,Anupam Basu, 2004). The main idea of the HMM is to find the sequence of tags for a given sequence of words. This can be done by combining the most frequent tag and the N-gram model that considers the tag sequence probabilities i.e. considering the lexical category of a word using its most frequent tag and its local context in the sentence from the training corpus during the training phase. Generally speaking, from a statistical point of view, the task of the HMM model is to find the most likely POS sequence T1, T2, … Tn for a given word sequence W1, W2, …Wn . In other words, the model has to maximize the conditional probability P() of the tag sequence given the word sequence over all possible tag sequences Tn (D. Jurafky ,J. H. Martin, 2006). Putting it all together, this approach tries to maximize the probability P(T1, T2, ...,Tn …… Tn | W1, W2, ...,Wn) which can be achieved by the help of N-gram model and most frequent tag model.

Generally, the positive characteristics of the stochastic method are researchers may not need language specialists; expertise coverage depends on the training data. The

disadvantages are it requires large amount of annotated training data (very large corpora) , some changes may require re-annotation of the entire training corpus in the supervised statistical learning, not easy to work with ill-formed input as both well-formed and ill-formed are still probable.

2.1.2.1. Hidden Markov Model

The most common model for stochastic approach is the Hidden Markov Model (HMM). It is the probabilistic function of Markov Process, a process which can move from state to state, from left to right on the states, to find optimal state sequence. AHMM is characterized by the following criteria (Christopher D., Manning Hinrich , 2000): Transitions among the states are governed by a set of probabilities called transition probabilities, a finite set of states each of which is associated with a probability distribution and in a particular state an outcome or observation can be generated according to the associated probability distribution. The observation is visible and the states are hidden to the observer and hence the name Hidden Markov Model.

HMM is defined formally as a set {S, O, A, B, } where (Christopher D., Manning Hinrich , 2000);

S denotes the set of states

O denotes the set of observation symbols.

A = {aij} is a set of state transition probabilities represented in transition probability matrix in which each aij represents a probability of moving from state Si at time t into state Sj at time t+1.

The state transition probabilities can be defined as aij = P(Si+1 = j|Si = i) for 1<=i<=n where n is the total number of states, aij>=0 and

B = bj(k) is an emission or observation probability distribution in each of the state S. bj(k) is the observation probability of observation k at the jth state.

The emission/observation probabilities bj(k) can be computed as bj(k) = P(Oi = k |Si =j) for 1<=j<=n and 1<=k<=m. bj(k) is the probability of state j taking the symbol Oi and it should be greater or equal to zero.

The initial state distribution$\pi = \{\pi i\}$ which is the probability of the first observation at a given state Si. Generally an HMM is the set containing {S, O,λ } where:

S = {S1 S2 S3 S4,....., Sn)

O = {O1 O2 O3 O4,....., Om)

λ = {A, B, π }

HMM takes three assumptions into consideration. The first assumption is called Markov assumption that states the first order transition probability can be extended to the kth order transition probability. It implies that, if it is possible to find the probability of state Si given the previous state Si-1 (P(Si|Si-1)), it would be also possible to find the probability of state Si given the previous K states (P(Si|S1, S2, ….Sk) (Jedlink, 2019). The second assumption is called the stationary assumption that says: state transition probabilities are independent of time and the output. It takes the due consideration of the actual time at which the state transition takes place and the output symbol that can be emitted being on the particular state (Jedlink, 2019). As a result, it considers that state transition probabilities are independent of the actual time and output symbol, which can be represented mathematically as:

P(St1+1|St1) = P(St2+1|St2) for any time t1 and t2.

The third assumption is the output independence assumption that states: the current observation is statistically independent of the previous observations. This means the probability of observing an output symbol Oi being in state Si is independent of the probability of observing Oi-1 being in state Si. This can be represented mathematically as P(Oi|S1, S2, S3…Sn). Because of these assumptions, HMM fails to accurately find the most likely sequence of states for a given sequence of observations.

When the HMM model is taken to the application of POS Tagging, the hidden states are the POS tags (tag sets) and the sequences of words are the sequence of observations. The transition probability in POS tagging is the probability of moving from one tag to the next tag and the emission probability is the probability of getting a word Wi being in tag Ti. The main problem of the HMM from the POS tagging point of view is, finding the sequence of tags thereby maximizing their probabilities given a sequence of words in a

text. This can be done using the Viterbi Algorithm that will take the joint probabilities of the lexical probability and the n-gram probability (Holen, 2009).

2.1.2.2. Artificial Neural Networks (ANN)

According to (Richard P, Lippmann, 1988), a neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes. In addition, according to (Haykin, 1994)A neural network is a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects:

1. Knowledge is acquired by the network through a learning process.
2. Interneuron connection strengths known as synaptic weights are used to store the knowledge.

Though the field of Artificial Neural Network was established before the advent of computers, the ANN simulations appear to be a recent development. This implies that the information processing of ANN was known in biological nervous system before actually applying it in information processing using computers applications. Hence, the basic idea of ANN here is to process information in a similar way that the biological nervous system process pattern. The key element of the ANN information processing scheme is the novel structure of the information processing system which is composed of, like the biological nervous system, a large number of highly interconnected elements called neurons working in union to solve a specific problem (Mohammed, 2010). Artificial Neural Networks learn from example by configuring for a specific application such as pattern recognition or data classification.

The ANN can learn by adapting different behavior on the basis of the data that is given to the network. It is possible to call the ANN learning an adaptive learning as the network is able to find properties from the presented data. It is not necessary to tell the network how to react to each data input separately like the conventional programming.

The common types of ANNs consist of three layers of units namely a layer of input units, a layer of hidden unites and a layer of output unites (Parikh, 2009). The input layer which is connected to the hidden layer represents the raw information that is fed to the network

as an input so that it can learn and adapt properties. The middle layer, so called hidden layer, connected to the output layer is determined by the activities of the input unit and the weights on the connections between the input and hidden units. The output layer represents the result of the learning properties from the input layer and hidden layer.

Taking the ANN approach to the application of part of speech tagging, first preprocessing activities are performed before dealing actually with the ANN based part of speech tagger. Such preprocessing activities can be tokenizing, feature extraction like POS information, word information, POS category and order information etc. The results of the preprocessing activities are given to the input layer of the network from which the network can learn pattern. As mentioned above, the input layer is connected to the hidden layer and in this layer different algorithms like error back-propagation algorithm, an algorithm based on an error-correction learning rule specifically on the minimization of the mean squared error which is a measure of the difference between the actual and the desired output, can be used for training the system (Parikh, 2009).

This technique of tackling the problem of assigning part of speech tags to words has some disadvantages compared to the HMM and rule based approaches. Some of these are: The HMM method assigns the sequence of tags for the sequence of words in the entire sentence i.e. it takes the due consideration of the sentence structure while most ANNs take the word to be tagged only. The same thing is true with rule based approaches which tend to take the sentence structure and generally the linguistic patterns into consideration (Schmid H. , 1994).

2.1.2.3. Hybrid Approach

As its name implies, this approach takes the combination of either rule based approach and stochastic approach or ANN and rule based by taking the advantages from both approaches to improve the performance of the system. Works like (Daelemans, W., J. Zavrel, and P. Berck, 1996) have used the hybrid approaches (rule based + stochastic) as a result they have got better results than the corresponding uncombined approaches.

2.2. Application Areas of Part of speech Tagging

Some of the applications of Part of speech tagging are**:**

Parsing: Parsing is used to assign syntactic description of the sentence. Parsers are useful for text analysis, corpora analysis, machine translation etc. The output of a POS tagger can be used as an input for syntactic analyzer or parser and the output of a parser in again can be used as an input for a semantic parser and automatic machine translation. The output of a tagger is used as input for a sentence parser, the performance of the parser will be improved. Since POS tagging assigns unique tag to each word this can help to reduce the number of parses. (James, 1995)

Information Extraction: Information extraction, on the other hand, is a means of retrieving certain types of information from natural language text automatically. The main aim of information extraction is to process natural language text and to retrieve occurrences of a particular class of objects or and occurrences of relationships among objects. Information extraction is also a form of natural language processing in which certain types of information must be recognized and extracted from text. It extracts their semantic contents tagging and helps to identify useful terms and relationships between them. In addition, POS references are used by patterns which are used for information extraction from text. POS tagging helps Information extraction to identify useful terms and relationships between the terms.

Information Retrieval: POS tagging provides information retrieval system with POS information which helps information retrieval system with more refined information so that information retrieval system can eliminate retrieval of irrelevant documents to the query.

Question Answering: POS tagging helps to analyze a query to what type of entity the user is

looking for and how this entity is related to other noun phrases mentioned in the question. The study of question answering systems, which enable people to locate the information they need directly from large free-text databases by using their queries, has become one of the important aspects of natural language processing and information retrieval and it is highly related to part POS tagging.

Speech Synthesis and Recognition: The most important information about the word and its neighbors (words that occur before and after the word) are useful in a language model for speech recognition. Such information can be derived from POS tagging. POS of a word can also indicate us something about how the word is pronounced depending on the grammatical category of the word.

Machine translation: POS tagging has also great influences on the probability of translation of word in source language to the word in target language in machine translation. POS tagging is the first and the very important task and step in machine translation and it can be part and component of machine translation system. In addition to above mentioned areas of application, POS tagging has also been used in other application like lexicography, word sense disambiguation etc.

2.3. Related work

The task of POS-tagging is attaching appropriate grammatical or morpho-syntactical category labels to every token, and even to punctuation marks, symbols, abbreviations, . . . etc. in a corpus. So POS is the first step in NLP that have different important application can be used it. Such as in machine translation, spell checking and correcting, speech recognition, information extraction, information retrieval, corpus analysis and text to speech synthesis system (HASAN MUAIDI, Levenberg-Marquardt, 2014). In this section, researches that are conducted on different approaches of POS tagging for different languages are reviewed. The purpose of this section is to see the development of different POS approaches and take on the crucial tools and implementations for the study.

Earlier part of speech tagging using neural network was done for Hindi language and compares them with two other machine learning approach, HMM and CRF (Ankur, 2009). In this research two network taggers are presented, the first one is single neural tagger is a neural network based POS tagger with fixed length of context chosen empirically is presented first. Then the second type of tagger is multi-neural tagger which consists of multiple single-neural taggers with fixed but different lengths of contexts are presented. Multi neural tagger performs tagging by voting on the output of all single neural tagger. A multi-layer perceptron network with three layers is used as a single

neural tagger and is trained in supervised manner with well-known error back propagation learning algorithm. Java programs have been implemented in order to prepare lexicon and implement MLP network and error back propagation learning algorithm and to achieve the results shown in this paper. In this research the corpus is divided in to three corpuses which are called development, training and testing corpus. The size of the corpus was divided into training; development and testing corpus were 187,095, 23,565 and 23,281 words respectively .Percentage of unknown words in the development and testing corpus were 5.33% and 8.15% respectively. Tools used for the experiment in this research Categories-MAP, a tool which finds stem, suffix and prefix from un-annotated text was used to handle unseen words. Brant's TnT HMM based tagger and CRF++, a CRF based tagger were used to compare performance of the presented taggers. The performance evaluations was done on both the development corpus and testing corpus for the three taggers called multi-layer tagger, HMM, CRF and have for HMM tagger the performance was 95.18%, 91.58% for the development and test corpus respectively. The Multi-neural tagger also has 95.78%, and 92.19% respectively. The CRF performance was resulted with 96.05%, 92.92% for the development and test corpus respectively.

A research work (HASAN MUAIDI,Levenberg-Marquardt, 2014) which has used Levenberg-Marqurdt learning neural network for POS tagging of Arabic sentences was done by Hassan Muaidi. In this research Levenberg-Marqurdt algorithm was used to train the ANN. Levenberg-Marquardt algorithm is an approximation to the Newton method used for training the neural network. In order to let understand the neural network and use the tokens, all letters should be converted or transformed to numeric value. In this research the binary coded method is used for the coding of the tag-sets. Input vector $\mathcal{X} = \langle X_1, X_2 \dots X_{12} \rangle$ with its corresponding output $\mathcal{Y} = \langle Y_1, Y_2 \dots \rangle$ are presented to the ANN which is considered as experimental data. Using the input vector$\mathcal{X}$, the output $\mathcal{O}$ is calculated this value is differs than the desired output $\mathcal{Y}$ and it is called the actual output or the net output. The difference between the desired and the actual output is computed and is called the error. After the error is calculated using the mean squared error (MSE) the error is propagated backward to change the weights in order to reduce or minimize the error rate. This process is repeated for a series of experiments until the error rates is

acceptable. For this research a corpora of 24,810 words are collected and manually tagged to train the neural network and to test the performance of the developed POS-tagger. The tag-set used in this research was adopted the tag-sets done by other researcher that contains a total 161 detailed tags and 28 general tags covering Arabic main POS classes and sub-classes and the Arabic letters are coded to a numeric value. The developed tagger achieved an accuracy of 98.83% when evaluated on the train set and 90.21% on the test set. And also the algorithm also compared with the existing Arabic algorithm called Back-propagation algorithm so it has better performance and efficient than the existing algorithm called Back-propagation algorithm.

A neural network based part of speech tagging for Hindi was done (Ravi Narayan, S. Chakraverty, V. P. Singh., march 2014). In this research two steps are done to complete the part of speech tagging; the first step is that the Hindi text are tagged using the rule based approach. Afterwards the second step, the tagged texts are checked for detection and correction of any anomaly using the neural network approach. The proposed tagger is work first raw sentences are passes through the tokenize system. In this step it splits the sentences into words and indexed it as token. The resulting words with tokens pass through rule based POS, simply using the lexicon. For correction and accuracy it finally passes through the ANN based POS tagger using the pattern recognition of corpus. To analyze the effectiveness of the proposed approach, 2600 sentences of news items having 11500 words from various newspapers have been evaluated. The POS tags in the sentences are represented using a numeric value. In this research uses binary representation for the POS tag. During simulations and evaluation, the accuracy up to 91.30% is achieved, which is significantly better in comparison to other existing approaches for Hindi parts of speech tagging.

The first Amharic part of speech tagger was done using hybrid (neural network and rule based) approach by Solomon Asres (Solomon A. , 2008). The researcher has conducted two steps to accomplish a better performance of tagger. The first step is that the Amharic text is tagged using the Neural Network approach. Afterwards, the second step is, the tagged text is checked for detection and correction of any anomaly using the rule based approach. He has adapted a multilayer perception neural network with back-propagation

algorithm and transformation based learning method for the development of Amharic tagger. He has used 30 tag-sets and 210,000 words of text corpus. He has used both lexical probability and contextual probability to find the most probable tag of a word. He has used both the lexical and contextual probability to find the most probable tag of word. The lexical probability is simply the probability of a word occurrence with a specific tag $\left(P(\mathcal{T}_i|\mathcal{W}_i)\right)$ that can be calculated by dividing the occurrence of the number of appearances of the $\mathcal{W}_i$ and $\mathcal{T}_i$ by the number of occurrences of $\mathcal{W}_i$ in the Text corpus. Contextual probability is the transition probability that can be determined by calculating the probability that the tag occurs with n-previous tags. The researcher has collected data from different source such as the Ethiopian Language Research center (ELRC), Addis Ababa. He has collected around 210,000 words from one ELRC project called "The Annotation of Amharic News Document" which is meant to tag each Amharic word in its context with the most appropriate part of speech manually. The project in turn has collected the sentences from Walta Information Center, a private News Agency located in Addis Ababa, Ethiopia, that makes daily news in Amharic and English through its website. To evaluate the proposed method, the researcher has conducted a lot of experiments. Relatively a large number of data is used to train and test the proposed tagger. As a result, the experimental performance of this work indicates that 91% and 94% accuracy for rule-based and neural network tagger, respectively. But the result reaches to 98% when the experiment has been conducted on the hybrid tagger. Though the text corpus taken for this thesis work is not as large size as that of brown corpus etc., it has achieved a higher performance on the hybrid approach.

According to (Adafre, 2005) part of speech tagging for Amharic using conditional random fields was done by implying two tasks the segmentation and part of speech tagging are carried out independently. This aim is to explore recent development in the morphological analysis of related languages, such as Arabic, Hebrew and machine learning approaches and apply them to the Amharic language. The tasks of Amharic word segmentation and part of speech tagging is performed using a small annotated corpus of 1000 words. The given the size of the data and the large number of unknown words in the test corpus (80%), an accuracy of 84% for Amharic word segmentation and 74% for POS tagging.

Another work for Amharic (Bjorn Gambackand, Fredrik Olsson, Atelach Alemu Argaw, Lars Asker) was done by applying the art of part of speech tagging to Amharic using three different tag sets. The taggers are HMM model (TnT), SVMTool, Maximum Entropy (MALLET). The datasets or corpus consists of all 1065 news texts (210,000) words are used from different resources. It has been morphologically analyzed and manually partof-speech tagged by staff at ELRC, the Ethiopian Languages Research Center at Addis Ababa University. . The best results were obtained using a Maximum Entropy approach, while HMM-based and SVM based taggers got comparable results.

According to (Martha Yifiru, Tachbelie Solomon, Teferra Abate , Laurent Besacier) is conducted to identify the best method for under resourced and morphologically rich language especially in case of Amharic language. In this research segmentation and tag hypothesis combination have conducted to improve tagging accuracy. Different POS taggers are used for the experiments, Disambig, Moses, CRF++, SVMTool, and MBT (memory based POS tagger generator) and TnT Trigram 'n' Tags. The Pos tag-set and the corpus used tokens are 210,000 developed within the AAND at the ELRC has been used for the experiment purpose. The larger the corpus and the higher the accuracy of the training set, the better performance of the tagger. TnT and SVM are affected by the amount of data used in the training and MBT is less affected by the amount of data used in the training. The result justifies that TnT works better with large training data size and MBT is less affected by the size of the data used in the training data set and also segmenting words which are composed of morphemes of different POS and which are assigned compound tags is a mean of improving tagging accuracy for the under resourced and morphologically rich languages. MBT and SVM based taggers have high performance for unknown words depending on the amount of data used in the training data set. The best accuracy was obtained using a Maximum Entropy approach when allowed to create its own folds: 90.1% on a 30 tag tag-set, and 94.6 resp. 94.5% on two reduced sets (11 resp. 10 tags), outperforming an HMM based(TnT) and an SVM-based (SVMTool) tagger.

Tigrigna part of speech tagger was done using hybrid (Hidden Markov model and rule based) approaches are done by Teklay Gebregziabher (Teklay, 2010) . The hybrid

approach was done by combining the hidden Markov model and the rule based approaches. He has used about 24,000 words from around 1000 sentences containing 8000 distinct words were tagged for training and testing purpose. The POS tag-sets used in this research are 36 tag-sets are identified for the tagger to put the appropriate word class. In this research first raw Tigrigna data are tagged using HMM tagger and afterwards using the rule based tagger will be corrected. Viterbi algorithm and Brill Transformation-based Error driven learning are adapted for the HMM and Rule based taggers respectively. He has used both the lexical and contextual probability to find the most probable tag of word. The lexical probability is simply the probability of a word occurrence with a specific tag $\left(P(\mathcal{T}_i|\mathcal{W}_i)\right)$ that can be calculated by dividing the occurrence of the number of appearances of the $\mathcal{W}_i$ and $\mathcal{T}_i$ by the number of occurrences of $\mathcal{W}_i$ in the Text corpus. Contextual probability is the transition probability that can be determined by calculating the probability that the tag occurs with n-previous tags. An experiment of the tagger is done on the 24,000 words for the training and testing purpose. From this amount of word 25% of the corpus was used for the testing purpose. As a result, the experimental performance of this work indicates that 89.13% and 91.8% accuracy for HMM and rule-based tagger, respectively. But the result reaches to 95.88% when the experiment has been conducted on the hybrid tagger.

Like ways, Tigregna Part of speech tagger by Mulugeta Atsbaha Sium, (SIUM, 2016) that uses a corpus containing 3100 sentences, 10000 distinct words and 56,151 total tokens and they are balanced corpus (not a domain specific corpus). A total of 22 Morpho-Syntactic course-grained tag-sets were adapted to prepare the annotated corpus using semi-supervised approach. Rule based, averaged perceptron taggers, and hybrid of the two taggers are investigated. The hybrid tagger was constructed from the sequence of the two taggers as averaged perceptron tagger followed by rule based tagger. The models are trained in 75% of the corpus and tested on the remaining 25% for their robustness and effectiveness. For each model several different experiments have been conducted. Experimental result shows that reasonable tagger is achieved with modified rule based tagger along to three combined initial state annotator. In this study state-of-

the-art tagging accuracy for morphological rich languages particularly Tigrigna with Averaged perceptron tagger is achieved.

The Rule based tagger has found 94.8%, while averaged perceptron tagger achieved 95.5%. Thus, averaged perceptron tagger and rule based tagger achieved comparable performance; however, the hybrid tagger improves the accuracy to 96.3%. The hybrid tagger works as a sequence of averaged perceptron followed with rule based tagger as error detection and correction sequence. In between the trained averaged perceptron and rule based tagger there is output analyzer with a threshold value as output validation and decision maker.

Another earlier work on Afaan Oromo language part of speech tagger was done by Getachew Mamo in Addis Ababa University in 2009 (Getachew M. , 2009). In this work, the researcher has used one of the known models, which is the generalization of the stochastic approaches, HMM for tagging Afaan Oromo Texts. He has collected 159 Afaan Oromo sentences (with 1621 distinct words) from different sources and he has used 17 tag-sets to annotate these sentences. He has divided these sentences as training set and test set. The HMM based Afaan Oromo part of speech tagger was trained on the training set in order to compute and store the lexical and contextual probabilities of words in the training. The tagger then takes untagged Afaan Oromo text as an input and tokenizes the sentences into words before actually assigns the part of speech tags sequence. After this, each token in the sentence is assigned with a correct part of speech tag sequence that is done using unigram and bigram models of the Viterbi algorithm by taking the knowledge from lexical and contextual probabilities gained during the training session. The researcher has tested the performance of the tagger by conducting experiments and as a result he has got an accuracy of 87.58% and 91.97% for the unigram and bigram models respectively.

There are also other works on Affan Oromo POS tagging by MOHAMMED-HUSSEN ABUBEKER, (Hussen, 2010),the study customized Brill transformational error driven learning tagger for Afaan Oromo. Some template in the original Brill tagger was modified to fit Afaan Oromo morphological nature. After training data is analyzed for its appropriateness using learning curve analysis, the study used 10- fold validation method for the experiment. Moreover experiment was conducted to determine the percentage of

training data for contextual and lexical rule learner. Best accuracy of the tagger was achieved when contextual rule learner training data is 35% and lexical rule learning data is 65%.This shows the morphological rule dominance over contextual rule for the language.

After modification on the templates of the Brill's tagger about 2.44% improvements over the original Brill tagger was achieved. This means 80.08% accuracy of the tagger was achieved in modifying the templates where the accuracy of the original tagger is 77.64%. Error of the modified tagger was also analyzed for further improvements using confusion matrix for the tagger. The result obtained in both original Brill tagger and modified Brill tagger is compared with Hidden Markov Model approach (bigram and unigram approach).The comparison shows that Brill tagger is by far better than Hidden Markov Model in all the cases for Afaan Oromo i.e Hidden Markov Model accuracy for bigram approach is 70.63% and for unigram 68.08% whereas that of original Brill tagger without modification is 77.64 and 80.08% for modified Brill tagger.

One of the latest research work on Affan Oromo langage by Getachew Emiru, (EMIRU, 2016)In this thesis, the development of part of speech tagger using hybrid approach that combines rule based and HMM approaches was conducted for Afaan Oromoo. The transformation based learner, which is a rule based tagger, tag the words based on rules, or transformations induced directly from the training corpus without human intervention or expert knowledge.

The HMM tagger, tags the words based on the most probable path for a given sequence of words. The hybrid approach of Afaan Oromo part of speech taggers developed in this thesis uses HMM tagger as initial annotators and Brill's' tag-ger as a corrector based on fixed threshold value. NLTK 3.0.2 and python 3.4.3 were used for the implementation and experiment. To minimize data requirement and the cost of data preparation they have used bootstrapping method.

To train and test the model 1517 sentences were used, that is collected from Afaan Oromo news agencies and Medias. For experimental analysis they have used 85% for training and the remaining 15% was used for testing. The performance analysis of the

three taggers, namely: HMM, rule based and hybrid tagger were tested with the same training and testing set they achieved accuracy of 91.9%, 96.4% and 98.3%, respectively. In conclusion, the accuracy of the hybrid tagger clearly shows that a clear improvement performance rather than separated taggers. To increase the performance of the tagger wide coverage/domain area of training data and morphologically segmented words were recommended for future works.

A research work on Wolaita language POS tagger by Berhanu Herano Ganta (Berhanu H. , 2015). The models or taggers were developed based on the review and the study made on the Wolatia language word classes. The tags developed for this study were the first attempt for the language and are based on the study and review of word class of Wolaita language. In this research, 200 sentences were manually tagged and of these 200 sentences, 90% of the sentences (180 sentences) were used for training and the rest of the sentences were used for testing the performance of the tagger. The result of this experiment showed that HMM based taggers perform better than CRF based taggers. The performance that have been achieved has accuracy of 83.58% and 74.63% using reduced tag set for supervised Hidden Markov Model (HMM) and Conditional Random fields(CRF) based taggers respectively.

In relation to Awngi language, (Tsegaye, 2013) did a research on developing a Stemming Algorithm for Awngi Text using longest match approach supplemented by context-sensitive and recoding matching principle. The stemmer is evaluated on Awngi text from three domains; news articles, text books and a dictionary. According to the evaluation of the stemmer, it is concluded that an overall accuracy of 91.41% is achieved which is a very good result as it is the first attempt to develop the algorithm. As the stemmer is the first kind for Awngi language, 8.59 % error is a number that can be minimized by introducing more rules and exceptional rules. Further research is not only required in the algorithm but also in the morphological structure of Awngi language.

CHAPTER THREE: METHODOLOGY

3.1. The Awngi Morphology

Morphology is the study of the structure of words, and of the way in which their structure reflects their relation to other words: both within some larger construction such as sentence and across the total vocabulary of the language. (Anderson, 1988). Each word consists of morphemes which are the smallest unit of a language that cannot be segmented further and contains a constant meaning. Root, suffix, prefix and ending or flexion are the basic types of morphemes.

To understand the morphology of Awngi we have, reviewed different sources such as textbooks, journals, articles and news papers. Additional information was also collected through personal discussion with professional experts of the language.

3.1.1. Noun morphology

A. Gender

As shown in table 3.1 below, Awngi has two genders; masculine and feminine. Masculine gender is marked by final /-i/ or a zero morpheme/ Ø /. The feminine is indicated by the ending/-a/. Most nouns referring to objects are masculine while use of feminine gender for these objects has diminutive or derogatory connotation.

No	Masculine	Feminine
1.	ግሴቸ Giseŋ-Ø ‘dog’	ግሴቻ Giseŋ-a ‘ bitch’
2.	ፊሪሲ Firisi ‘horse’	ፊሪሳ Firisa ‘ mare’

Table 3.1 inflection of gender

B. Number

Awngi has two numbers; singular and plural. There is no gender distinction in the plural. The most plural marker is /ካ፧-ka/. It comes right after the stem.

Table 3.2 shows plural nouns with /ካ፣-ka/.

No	Singular	Plural
1.	ታይ Tay ‘ram/sheep’	ታይካ Tay-ka ‘sheep’
2.	ፊሪሲ Firisi ‘horse’ ፊሪሳ Firisa ‘ mare’	ፊሪስካ Firis-ka ‘ horses’
3.	ልኩ likw ‘leg’	ልኩካ likw -ka ‘legs’
4.	እንፁ Instu ‘thin’	እንፁካ Instu-ka ‘thin’

Table 3.2 inflection of number

As we can see from the gender markers /-i/ and /-a/ are deleted before plural marker /-ka/. In the words, /-ka/ is attached to the stem or base form. Awngi has some nouns which reduplicate for plural, as ስር sir ‘child’, ስራስሪ sarasiri ‘children’ , ኪሲ kisi ‘priest’ , ኪሳኪሲ kisakisi ‘priests’, ኹና xuna ‘woman’ , ኹናኹና xunaxun, women. All adjectives modifying plural nouns take plural marker suffix/-ka/, with rare exceptions where base forms may optionally be used as plural as shown below.

ሻሳነት-ካ አቃ ይነቱና
Xisant-ka aq yintuna
Elderly people come

3.1.2. Morphology of Personal pronouns

The subjective forms are /አን *an* /and /እንት *int* /while the oblique forms are /ይ *yi-/* and /ኪ *ki-/* respectively. The later appears only bearing case marking morphemes; they do not appear independent of the markers. Plural pronouns suffix /ስ -s /before they show accusative commutative-directional and genitive cases. The singular of second and third persons have reverent forms when used to express respect. These are: /እንቱ into/ and /ቻ ŋa / respectively.

3.1.3. Cases

Case is the syntactic feature that defines relation between noun phrases and their predicates and it applies to all phonetically overt noun phrases. Case is the system of marking dependent nouns for the type of relation they bear with their heads. The verb is taken to be the head of the clause since it largely determines what dependants may be present.

A. Accusative case

Accusative case in Awngi is expressed by inflectional elements and suprasegmental feature. The inflectional elements are/-o/,/-wa/,/-e/ and /-sa/.

Let us see the following examples

A) ግሴቻዋ ኩና

Giseŋa-wa ku-na

They killed the bitch

In the above example (A), the accusative case marker/-wa/ is suffixed to Giseŋa 'bitch' to mark case. The occurrence of /-o/, /-wa/ and/-e/ is phonologically conditioned. /-o/ occurs after consonants or/u/; /-wa/ occurs following/a/ or/-e/ and /-e/ occurs with nouns ending in /i/ which deletes after affixation of [-e].

B) ግሴኝ ኩና

Giseŋ-o ku-na

They killed the dog

C) ግሴኝ-ካ-ዋ ኩና

Giseŋ-ka-wa kuna

They killed the dogs

D) ዛግሬ ኩና

Zagri +-e zagr-e ku-na

They killed the monkey

As can be seen from the examples (B, C, D), the accusative marker is/-o/ in nouns that end in consonant, /-wa/ that end in a vowel/-a/ and /-e/ in nous that end in /-i/. The occurrence of /-sa/ is grammatically conditioned in that they occur with possessive forms that serve as qualifiers and with verbs of relative clauses. The following examples (E, F) shows the accusative case marker for/-sa/

E) ይውሳ ፍያሎ ኩና

Yi-w-sa fiyal-o ku-na

They killed my goat

F) ይትሳ ፍያላዋ ኩና

Yit-sa fiyala-wa ku-na

They killed my female goat

B. Dative case

Dative is typically the case of indirect object. In Awngi, dative case is marked by the suffix/-s/ that occurs attached to the indirect object, which comes following the direct object and before the verb. Consider the following examples (G, H)

G) ገንዜቦ ኹናስ ይትኾማ Genzab-o xuna-s yitixwa-ma

Did u give the money to the woman?

H) ገንዜቦ ፉቻሰ ኹናስ ይትኾማ

Genzab-o fuca-s xuna-s yitixwa-ma

Did u give the money to the white woman?

In (G), The dative marker (-s) is suffixed to the head noun xuna ‘women’ but in (H) it is suffixed to both the qualifier and the head which is not common.

In (G), The dative marker (-s) is suffixed to the head noun xuna ‘women’ but in (H) it is suffixed to both the qualifier and the head which is not common.

C. Comitative Case

Comitative case indicates the accompaniment in action. In Awngi, the suffix /-li/ is expressed this case.

Consider the following examples (I, J)

I) ታብሊ ጀርሊ ይንትኺ

Tabli jer-li yintixwa

The father came with his son

J) ፉቻ ኹናሊ ካስኾ

Fuca xunna-li kasixwa

He went with the white woman as shown in (I, J), /-li/ is suffixed to the noun jer and xuna.

D. Genitive case

Blake describes the genitive case as encoding the abdominal relation that subsumes the role of possessor, and the label possessive case as a common alternative. The occurrence of Awngi genitive marker is very complex. They vary in accordance with not only the number and gender of the possessed nouns but also for phonological reasons. The possession suffixes are /-u/,/-t/,/-ti/,/- ku/, /-kw/ and /-su/.

/-u/ when the possessed noun is singular masculine.

/-t,-ti/ when the possessed pronoun is singular feminine.

/-ku,kw/ when the possessed noun is plural

/-su/ after the plural pronoun but only when the possessed is singular.

Consider the following examples (K, L, M, N, O, and P)

K) አቓው ቢሪ

Aqqa +u Aqqa-w biri

The woman's ox

L) አቒት እሏ

Aqqi-t illwa

The man's cow.

M) Aq-ti illwa

The men's cow.

N) አቓካው ፈየልካ

Aqqa-kw fiyala –ka

The woman's goats

O) አቕኩ ፍያልካ

Aq-ku fiyala –ka

The men's goats

P) እኖጂሱ ኻን

Innji-su ŋin

Our house

The phonological conditioning concerns /-u/ and/-w/, /-t/and/-ti/ and /-ku/ and/-kw/. /-t/and /-kw/ occur following a vowel whereas /-ti/ and/-ku/ occur following a consonant. The occurrence of suffix /-su/ is limited to plural pronouns.

E. Local cases

Local cases express notations of location ('at'), destination ('to'), source ('from') and path ('through') . The distinguishable local cases in Awngi are locative, ablative, directional, directional-comitative, and purposive. Locative case shows location and shown by suffix /-da/ which is suffixed to a word referring to a noun. Ablative case express the origin and is shown by suffix /-das/. Directional is self explanatory and is expressed by suffix/-so/. Directional comitative shows not only direction but also

accompaniment in that something or someone has gone or been taken to some other person to live or stay with them and is shown by suffixes /-wla/,/-ula/ and /-sula/.

Consider the following examples (Q, R , S, T, U)

Q) ኹና ŋ̈ንዳ ዝኬ

Xuna ŋ̈in-da zike

The woman is in the house

R) ኹና ŋ̈ንዳስ ቲንትኳ

Xuna ŋ̈in-das tintixwa

The woman came from the house

S) ኹና ŋ̈ንሶ ካትኾ

Xuna ŋ̈in-so katixwa

The woman went towards the house

T) ኺ ታላኡላ ካታ

Ŋ̈i-tala-sula ka-ta

She has gone to her father

U) ብና ካታ

Bin-a ka-ta

She has gone to river (for fetching water of washing clothes).

3.1.4. Verb morphology

In Awngi, there are two clearly identified aspect of a verb; perfect and imperfect. As aspect is a term that covers how we view an event within a time frame. The perfect aspect refers to temporally bounded situation whereas imperfect aspect refers to a situation which is not temporarily bounded but which is rather an explicit reference to its internal structure. Hence, perfect aspect is very often related to past tense while imperfect aspect refers to habitual, continuity, progressivity, and etc. table 3.3 shows the perfect and imperfect forms of the verb ሱግ sug ‘pound’.

Person			Verb aspect		Aspectual suffixes		
			Perfect	imperfect	perfective	imperfective	Jussive
1st	Singular		Sugixwa	Sug-a	- ixwa	-á	Sug-is
	Plural		Sug-nixwa	Sug-n-a	ixwa	á	Sug-nis
2nd	Singular		Sug-tixwa	Sug-t-a	ixwa	- á	Sug-tis
	Plural		Sug-tun- a	Sug-t-a	-a	- á	Sug-tin-Is
3rd	Singular	Masculine	Sugixwa	Sug-a	- ixwa	- á	Sug-is
		Feminine	Sug-tixwa	Sug-t-a	- ixwa	- á	Sug-tis
	Plural		Sugun-a	Sug-an-a	- á	- á	Sug-inis

Table 3.3. Awngi Verb morphology

3.1.5. Morphology of numerals

Numerals are those items that include the cardinals (one, two, and three...) and the ordinals (first, fifth…). The process of forming ordinals from the cardinals is as follows:

ላኹ - one እምፕላንቲ - first

ላቻ - two ላቻንቲ - second

ሹካ - three ሹካንቲ - third

ሴዛ - four ሴዛንቲ - fourth

እንኪ - five አንኪንቲ - fifth

ዋልታ - six ዋልታንተ- sixth

The morpheme of the ordinal marker is /አንቲ/ which is suffixed to cardinal numbers ending in a consonant and /ንቲ/ is suffixed to those that end in a vowel /አ/.

In general, the morphological structure of Awngi language shows that both the inflectional and derivational morphologies involve suffixing. Analysis of the Awngi language and grammatical structure revealed that words are created by adding suffixes to a basic stem. It is also shown that the language uses an extensive concatenation of suffixes. This characteristic of the language make a very short stem into a long word. As

it is evidenced by different authors, such nature of the language indicates the complexity of the language's morphology. The complexity of the language is one of the main reasons for conducting a research on a Part of speech tagging.

3.2. Awngi POS tag sets

Awngi is an official language of Awi zone which is used as a medium of instruction in elementary schools and it is given as a diploma program in Injibara Teachers Training College. It uses a special character representation called Ge'ez. The Awngi language has its own distinct way of grammar construction, character representation and sentence formation. Awi people have been granted their own Nationality Zone in the Amhara National Regional State and have decided to establish Awngi as the medium of instruction for primary and higher education. Therefore, orthography was created in the late 1990's. Using this orthography, some textbooks were published and are now used in primary schools.

3.2.1. Awngi sentence structure

One of the levels used in the analysis of natural language processing is lexical analysis. Lexical analysis is used to interpret the meaning of individual words and it is used for word level understanding. Part of speech tagging plays great role at this level by assigning single part of speech tag to each word and the most probable part of speech to words that have more than one part of speech based on the context in which the words occur ((GANTA, 2015).

Words are traditionally grouped into equivalence classes called parts of speech (word classes, morphological classes, lexical tags. In traditional grammars there were generally only a few parts of speech (noun, verb, adjective, preposition, adverb, conjunction, etc.) Whereas, currently there are many more word categories added for natural language processing in general and part-of-speech tagging in particular to identify words in sentences with their specific identities.

3.2.2. Word formation in Awngi

According to (Tsegaye, 2013), Awngi is mostly an inflective language and has nine general categories or classes of words.

I. nouns (e. g. አቿ - a man, ¦ሻን-a house);

ላፓ አቿ ቢሩሾ ቱና። /two men's entered into office.

II. pronouns (e. g. እንት - you, አን - I);

እንትስታ አን እምፕልዳ እንክርንስ/lets play together.

III. verbs (e. g. ትንክፍ - to push, አንቤብ - to read);

አንቤብሻ ዋኽ አቿዄዊ።/reading makes a full man.

IV. adjectives (e. g. አዊ - sunny, ድሚ - red, ሊጊሲሚ-tall);

አበበ ከበደዴስ ሌጌሴምቴ።/Abebe is as tall as Kebede.

ኢትዮጵያኻዳ ኽሳንቲካዌን ራስደጀን ያኽ።/Ras Dejen is the highest mountain in Ethiopia

V. numerals (e. g. **አንኩዋ**- five, ሻይ -a thousand);

VI. adverbs (e. g.አይፓ - yesterday, እንማቺ- nearly, ፓ- tomorrow);

.አይፓ አይሊኔ እሬ እሉኻ።/ There was heavy rain yesterday.

VII. prepositions (e. g. ሊ - with);

VIII. conjunctions (e. g. እስታ - and);

IX. particles (e. g. ይኹቺ - only, ሚንች- much፣);

3.2.2.1. Awngi noun class

A noun is a word that is used for labeling things, such as a real thing (for example, bird), an imaginary thing (for example, ghost), an idea (for example, love), person name (for example, "kebede"). Awngi nouns, like English, are words used to name or identify a class of things, people, places or ideas. They typically function as arguments, subjects, objects of transitive verbs or complements of prepositions.

Examples of noun

አቂ(xqi) Man

ልኻኽ(lehahi) Dark

ታይ(tay) sheep

ቻካ(caha) Bird

3.2.2.2. Awngi Pronoun class

3.2.2.2.1. Personal pronoun

Like nouns, Awngi pronouns inflect for cases except for nominative which is unmarked. The first and second person singular is suppletive for the oblique case. In the first and second persons of the singular, there are distinct forms for subject and oblique pronouns, but only one form for the rest. The subjective forms are /አን *an* /and /እንት*int* / while the oblique forms are /ይ*y*and /ኪ*ki-*/ respectively. The later appears only bearing case marking morphemes; they do not appear independent of the markers. Plural pronouns suffix /ስ -s /before they show accusative commutative-directional and genitive cases. (Tsegaye, 2013)

Category	Singular	Plural
1st person	አን(xne)(I)	እኖጂ(xenoji)We
2nd person	እንት(xente)You	እንቶጂ(xenetoji)You
3rd person	ኺ(gni) He/she	ጋጂ(gnagi) They

Table 3.4. Personal pronoun subject position personal pronouns

For example the following sentence illustrates the above personal pronouns

Example 1: **አን ክንታንቴኽ**።xnekenetanetyhe / I am a student.

እንት ጊዪኔኽ።xenetegitsinyhe/ You are a merchant.

ኺ አሬሳቻንቴኽ::gnixrysacanetehe /He is a farmer.

እኖጂ ውታድርካኽ::xenojiwetaderekahe / We are soldiers.

3.2.2.2.2. Possisive personal pronouns of Awngi

Possessive pronouns are pronouns that indicate ownership of something

Category	Singular	Plural
1st person	ይው (mine)	እኖጅሱ(ours)
2nd person	ኩው(your)	እንዮጂሱ(yours)
3rd person	ሺው(her/his)	ቻጂሱ(theirs)

Table 3.5. Awngi language possesive pronoun

3.2.2.3. Awngi verb

Awngi language has generally a subject, object, verb (SOV) word order. The Verb is a word that tells us the state of doing or being. Awngi verbs carry inflections of aspect and mood and hence are morphologically the most complex. A lot of words with other POS are derived primarily from verbs. There are two major approaches to identify verbs from other word categories: syntactical and morphological approach. In the former case, verbs function as predicates in a simple sentence and they are found at the end of a sentence. In the later case, they reflect grammatical categories such as aspect, mood and agreement.

Examples ዱንፀኙ / break

ዘሚርኙ/ song

ካፅኙ /took

3.2.2.4. Awngi Adjectives

Terms or words that clarify nouns are known as adjectives. Awngi adjectives are words used to express things of behavior, shape, quantity, color, e.t.c

Examples of Some of Awngi adjectives

To express Behavior	To express Quantity	To express Shape	To express Color	To express clan/factin
ሸጊ(segi)Good **ድኪ**(deki) Bad **ጊሪቲ** (giriti)Humle	**ጽሊ**(teli)Samll **ድንጉሊ** (denuli)Big **ሚንቺ** (minei)Much	**ዋኽ**(wake) Circle **ሙላሊ** (mulali)Oval	**ድሚ**(demi) Red **ፃርኪ** (tarki)Black **ሞሹ**(mosu) Yellow	**ኢትዮጵያዳ** Ethiopian **አሚሪካዳ** American

Table 3.6. Awngi Adjectives

3.2.2.5. Awngi Numerals

Numerals are those items that include the cardinals (one, two, and three...) and the ordinals (first, second, fifth…).

The process of forming ordinals from the cardinals is as follows:

cardinal numbers	ordinal numbers
ላኹ(lahu) - one **ላጋ** (laga)- two **ሹኻ** (shuha)- three **ሴዛ**(seza)- four **እንኳ**(Ankua) - five **ዋልታ**(walta) - six	**እምፕላንቲ**(emplanti) - first **ላጋንቲ**(laganti) - second **ሹካንቲ**(shuhanti) - third **ሴዛንቲ**(sezanti) - fourth **አንኳንቲ**(ankuanti) - fifth **ዋልታንት**(waltanti)- sixth

Table 3.7. awngi numerals

The morpheme of the ordinal marker is /□□□/ which is suffixed to cardinal numbers ending in a consonant and /□□/ is suffixed to those that end in a vowel /□/.

3.2.2.6. Adverbs

Adverbs are any words that explain or modify verbs. These can be adverbs of time, place, manner, frequency etc.

some of the examples of Awngi Adverbs are expressed in the following table

Adverbs of time	Adverbs of place	Adverbs of frequency	Adverbs Manner
□□□ (segela) morning □□(hari) night □□(nura) Always □□□□(xenekogna) last year	□□□(xfeda) Outside □□(xhe) Inside □□□ (xeneda) This □□□ (xneda) That	□□ □□ □□□	□□□(dekegna) Badly □□□(watgna) How

Table 3.8. Awngi Adverbs

3.2.2.7. Conjunction

A word that can be used to join or connect two phrases, clauses and sentences is known as a conjunction. Conjunctions can be divided into coordinating and subordinating conjunctions. Coordinating conjunctions are used to connect two independent clauses. Mostly these conjunctions are used when the speaker needs to lay emphasis on the two sentences equally.

Here are some of Awngi conjunctions

እስታ (xeseta) and

ጂስታ (giseta) like

ያኹ(yahu) or

ያኻማኪ(yahamaki) therefore

3.2.2.8. Awngi punctuation Marks

All Awngi punctuation marks like **:-, :, ::, ?** and **!** are assigned the tag PUNCT.

3.3. Tagset for Awngi language

A sentence is composed of two components and these two components in turn consist of words. Therefore, it is possible to conclude that words are the basic components of every sentence. The meaning of a sentence is analyzed from the meaning of individual words and the way they are arranged. This shows that words are rarely used alone. Words more often work together in small groups which together make up whole sentence that possess

a single coherent meaning. Moreover, the same word can be used in different sentences and belong to a different world class category.

Three basic criteria are considered in order to categorize words in a language. They are: the meaning of the word, the form or shape of the word, and the position or the environment of the word in a sentence. These can be taken as the main criteria to determine the categories of a given word (Teklay, 2010)

As far as the researchers‟ knowledge is concerned, except that of Ethiopian languages i.e. Amharic, Affan Oromo, Tigregna, Wolaita, etc, for Awngi there is no readymade tagset , identified. This implies that identifying and developing tagsets for the language is mentioned to be paramount significant for this thesis work. As a result, the researcher has made continues discussion with Awngi Language professionals who have knowledge on the language. The preparation of tagsets for a language is tedious and time consuming task, which needs in fact, human experts in the language of interest, after continuous and intensive discussion with them, the researcher have developed broad category of tag sets. Moreover, the works of POS tagging on other Ethiopian languages are used as a reference.

Examples of awngi sentences with part of speech

ክንታኑቲ አይኝ ኮቢ ጂውኻ። / A student buy pen yesterday.				
ክንታኑቲ/nnprep	**አይኝ**/Adv	**ኮቢ**/nn	**ጂውኻ**/vv	**።**/punct

Table 3.9. Awngi sentences with part of speech tagging

Table 3.9. Awngi sentences with part of speech tagging

The tagsets that are discussed below are classified as a basic class and subclasses of the basic class where noun, pronoun, verb, adjective, preposition, conjunction, adverb, interjection are considered to be the basic classes. In addition, numeral and punctuation are also included as basic classes in the process of identifying the tagsets.

The concept hierarchy of the total tagset that are identified is presented in the following table.

No	Main category	Derived category/tags	Description	Example
1.	Noun	NN	A tag for all nouns	**አቺ**(man) **ወንበር**(chair) **ኻግ**(bed)
		NNNADJ	A tag for all nouns combined with adjective	**እሎ-ድሚ**(red eye) **ልኮ-ካጊ**(unlucky) **ፃርኪ-ታይ**(black ship) **ኬቴመዝኩዋንትካ** (city residents)
		NNCONJ	A tag for all noun combined with conjunctions	**ከበደስታ አለሙ** (Kebede and Alemu), **አበበያኹ ከበደ**(Abebe or Kebede), **ምራብሾ** (to the west)
		NNPREP	A tag for identifying noun combined with preposition	**ባህርዳርዳ** (from Bahirdar) **አለሙጂስታ** (Like Alemu)

Table 3.10. Awngi noun tagset identified

No	Main category	Derived category/ta	Description	Example

		gs		
2.	Pronoun	PR	For Personal pronoun	**አን**(me) **እኖጂ**(we) **እንት**(you) **ሻይ** (he/she) **ጋጂ**(they)
			For all demonstrative pronouns	**እን**(this) **ሻይ**(he/she) **አና**(those)
			For all possessive pronoun	**ይው**(mine) **ክው**(yours) **እኖጅሱ**(for us) **እንቶጅሱ**(for you all) **ጋጂሱ**(for them)
			For Interrogative pronoun	**አይ**(who) **እንዳራ**(what) **ወዳ**(where) **ወኒ**(when) **ዳማስ**(why) **አው.ይ**(whom) **ወታ.ያኻኒ**(how)
		PRCONJ	For pronoun + conjunction	**ይጋስታ**(like me) **እንትስታ አን** (you and me)
		PRPREP	For pronoun+ preposition	

Table 3.11. Awngi Pronoun tagset

No	Main category	Derived category/tags	Description	Example
3.	Verb	VV	A tag for all verbs	**ዱንፀ.ኻ**(Break) **ዝቾኾ**(drink) **ኹሩ.ኻ**(sleep) **ብብር.ኾ**(light) **አዜንኾ**(sad)
		VVPREP	A tag for identifying verb combined with preposition	**ዴማክሻስ**/to add **ካፃንታኪላ**/in order to take
		VVCONJ	A tag for identifying verb combined with conjunction	**ኽይስታንታ**(to release), **ፋይፃንቲስታ**(to require)

Table 3.12. Awngi verb tagset

No	Main category	Derived category/tags	Description	Example
4.	Adverb	ADV	A tag for adverb	አኽ(inside) ፅላ(small), አይኝ(yesterday), ድክኝ(badly)
		ADVNN	A tag for adverb and noun	ሚንቸጊዞ(Most of the time)
		ADVPREP	A tag for adverb and preposition	ንኽዊኒ/(this year), ኹሉቺፋሱ (always)

Table 3.13. Awngi Adverb tagset

No	Main category	Derived category/tags	Description	Example
5.	Adjective	ADJ	A tag for all adjectives	ድኪ(Bad), ድንጉሊ(big) ክቤብ(circle) ድማ(red) ፃርኪ(black) e.t.c
		ADJPREP	A tag for adjective and preposition	እንስኪላ(For this)

Table 3.14. Awngi Adjective tagset

No	Main category	Derived category/tags	Description	Example
6.	Preposition	Prep	Preposition	
7.	Conjunction	Conj	A tag for identifying Conjunction	ያኻማኪ(therefore), እስታ(and), ያኹ(or) ስ(for), ጃላ(on)
8.	Numerals	CD	A tag for all cardinal numbers	ላኹ(one),ላኝ (two) ሹኻ (three),ሴዛ (four)
		OD	A tag for all ordinal numbers	እምፕላንቲ(1ˢᵗ) ላኝንቲ(2ⁿᵈ) ሹኻንቲ (3ʳᵈ) ሴዛንቲ (4ᵗʰ)
9.	Punctuation	PUNCT	A tag for identify punctuation marks	:-, ፤,፣ ::, ?, !
10.	Interjection	INT	A tag for identifying emotions	አች(oh) አሃ(waw)

Table 3.15. Identified Awngi tag sets

As shown on the above tables, after we have explored the morphology of Awngi with experts, for tagging purpose we identified main tagset and derived tagsets. Words categorized under main tag sets are belongs to either of one of the part of speech tags. For example, In Awgni if words that end in ካ(ca) are likely to be nouns (NN). The word ታይ(sheep) and ታይካ(sheeps) are both nouns(NN).

Derived tag sets are used to tag the composition of more than one word together. As it has been discussed on the above sections, The Awngi language and grammatical structure revealed that words are created by adding suffixes to a basic stem. It is also shown that the language uses an extensive concatenation of suffixes. This characteristic of the language make a very short stem into a long word. So that, in order to identify such types of words we have used derived tagsets for example, the word ባህርዳርዳ(from Bahirdar) is categorized under derived tag sets of NNPREP(a tag for noun and preposition).

3.4. Data preparation

The tagged corpus is the immediate requirement for different analyses in the field of Natural Language Processing (NLP). Most of the language processing works like part of speech (POS) tagging is in need of such large collection of texts, which provide a real, natural, native language of varying types. Since there is no manually tagged corpus which is useful for NLP tasks like POS tagging for Awngi language, it demands to collect and prepare such a corpus to conduct POS tagging experiment. For the purpose of this research, data is collected from different sources like elementary and high school text books, from Amhara Mass media agency Awngi radio and television program, as well as from Awi zone administration office. In addition we have used the books which were prepared by Aysheshim Abate, 2018 and Gebre Bizuneh, 2018. The title of the book was **አውኺሰስታ አማኻርኺስ** which consists of 122 pages and **የአዊኛና አማርኛ ቋንቋዎች ንጽጽራዊ ትንተና** 156 pages respectively.

To conduct experiment in this research, only small sample of about 450 sentences and more than 3500 tokens was used due to there is no corpus annotated with part of speech for Awngi language and it is also time consuming, laborious and expensive to tag large amount of corpus manually.

The corpus prepared to conduct experiment was divided into two sets, training and test set. The training set was used to develop the model and the test set was used to evaluate the performance of the model. The training set consists of 90 percent of the total data set and the rest 10 percent was for testing set.

3.5. Evaluation Procedures

The taggers were trained on training set which was prepared in the previous stage. The result obtained on the training set was evaluated by comparing it with manually tagged corpus. The taggers or the models were then tested on the test set with untagged data. The test was used to see how well the models or the taggers perform on unseen data. Finally, the output of the tagger was compared with that of manually tagged data. The performance of the tagger was evaluated by dividing the number of correctly tagged words to the total number of words in the test set. The performance evaluation was done using tenfold cross validation which is used for evaluating of the performance of the tagger.

3.6. Approaches and algorithm design

Unlike stochastic approach, rule based approach requires frequent involvement of language experts in order to tag each token to the appropriate tag set. Among the different types of approaches which are listed on the previous chapters, statistical (stochastic) approaches are used. It is because most current part of speech taggers are probabilistic and is referred to tag for a word by calculating the most likely tag in the context of the word and its immediate neighbors. A stochastic approach includes most frequent tag, n-gram and hidden Markov model (Milion Meshesha, Getachew Mamo, 2011).

HMM is the statistical model which is mostly used in POS tagging which enables to estimate the most likely sequence of tags, making use of observed frequencies of words and tags in a training corpus.

For algorithm design and implementation, HMM is used for the study since it does not require detail linguistic knowledge unlike that of rule based tagger. From the different tagging algorithms viterbi algorithm is selected. Viterbi algorithm is a dynamic

programming algorithm that optimizes the tagging of a sequence, makes the tagging more efficient in both time and memory consumption. (simon, 2000)

3.7. Design of Awngi Pos Tagger

Awngi POS tagger is a program that assigns part of speech to words according to the context of that word in a sentence. As it is discussed in chapter one, statistical approach is used for this work, and it uses lexical and contextual rule to assign part of speech to a given word. This POS tagger first uses statistical techniques to extract information from the training corpus and then uses a unigram and bigram probability to automatically estimate the appropriate tag set for a given word in the corpus.

Assigning grammatical categories to words in a text is an important component of a natural language processing (NLP) system. Text collection tagged with Part of speech (POS) information are often used as a prerequisite for more complex NLP applications such as information extraction, syntactic parsing, machine translation or semantic field annotation etc. Awngi POS tagging is a method of assigning a specific part of speech tag to each word in a sentence to disambiguate the function of that word in the specific context. In this section, a detail description of design issues and techniques of the Awngi POS tagger are discussed.

3.7.1. Design goals

The main goal of designing HMM POS tagger is to achieve better performance in tagging Awngi texts. In addition to this, the tagger is expected to be easy to implement and increase speed in responding a tagged text, and easy in obtaining the required knowledge for the future researchers in this area.

3.7.2. Lexical and Contextual Probability in Hidden Markov Model

3.7.2.1. The lexical model

The goal of the lexical model is to prepare lexicon and the lexical probability of each word for each tag in the training set. The lexical probability can be calculated with relative frequencies using the following formula: $P(wi|ti) = \frac{\text{count of(Wi,Ti)}}{\text{count of (Ti)}}$

Where Wi and Ti are the i^{th} word in the input sentence and the i^{th} tag in the tagset respectively. The relative frequencies for the lexical model can be found by counting every word with a specific tag and divide it with the number of occurrences for this particular tag, which gives the conditional probability of the word given the tag (Teklay, 2010).

3.7.2.2. **Contextual model**

This model helps the HMM tagger gather the context of words in the training corpus as lexical model only deals with the probability of the word given the tag. I.e. relying only on the lexical model may degrade the performance of the tagger and hence it is important to take context of words into consideration. The contextual model also called N-gram Model that considers the sequence of part of speech tags is aimed to calculate the transitional probability of tags. The training data contains small training corpora; it would be convenient to use the N-gram model to be bigram or trigram model that considers the previous one or two tags respectively. Since this research work has small training corpora in comparison with the corpora for Amharic and brown corpus of English, a bigram model is selected. Therefore the contextual probability is found via tracking the previous one tag which can be calculated using relative frequencies by the following formula.

$$P(Ti|Ti-1) = \frac{\text{(Ti,Ti}-1)}{\text{(Ti}-1)}$$

The relative frequencies can be calculated by counting the frequency of Ti and Ti-1 and divide it by the number of occurrences of Ti-1 in the training corpus.

The HMM strives to find the optimal sequence of part of speech tags for a sequence of words in an input sentence using Viterbi algorithm. The tagger gets the lexical probability and contextual probability from the training corpus.

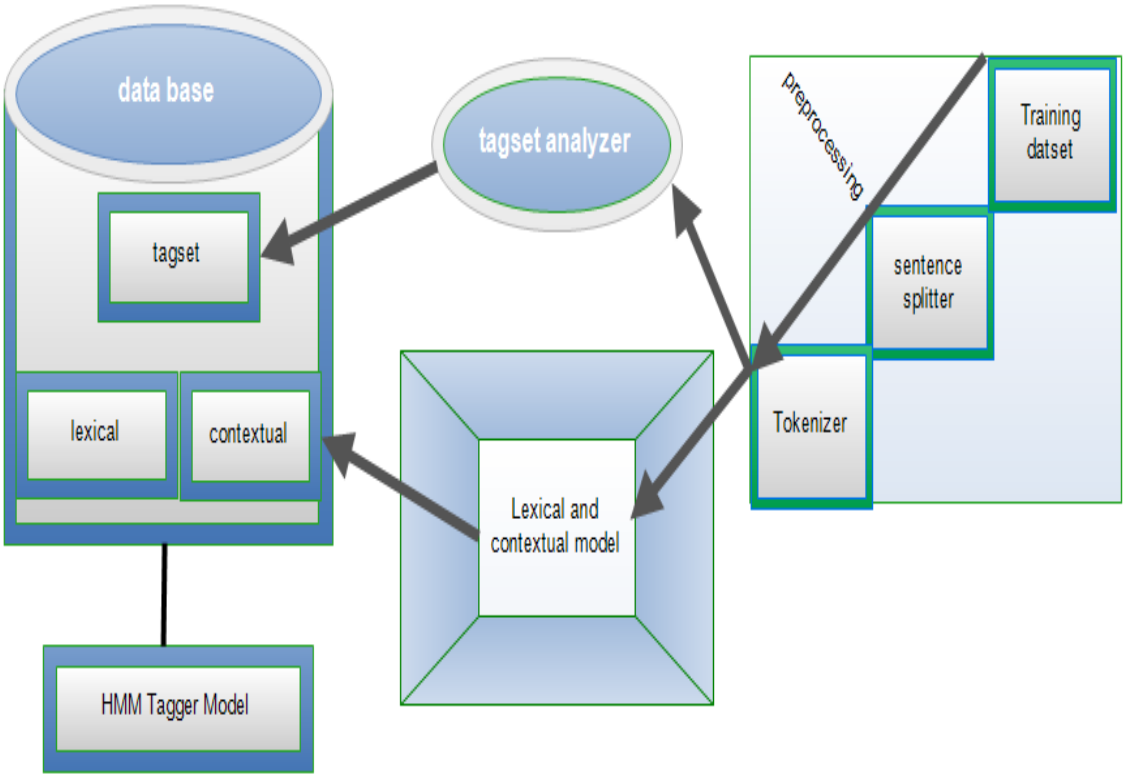


Figure 3.1. HMM tagger trainer model

The above figure 3.1 Shows the HMM tagger trainer model. A supervised learning method is used for training the HMM model. i.e. The training corpus is part of speech annotated Awngi text. The tagged corpus is an input to the model, and then it is given to the sentence splitter module in order to prepare it in a sentence level for training. The segmented sentences are given to the Tokenizer for splitting each sentence to a word level. After each sentence is tokenized into words, a tagset analyzer extracts the tags from the words and stores them in the database. The lexical and contextual models compute the lexical and contextual probabilities which are important for finding a sequence of part of speech tags for the sequence of words in the input sentence.

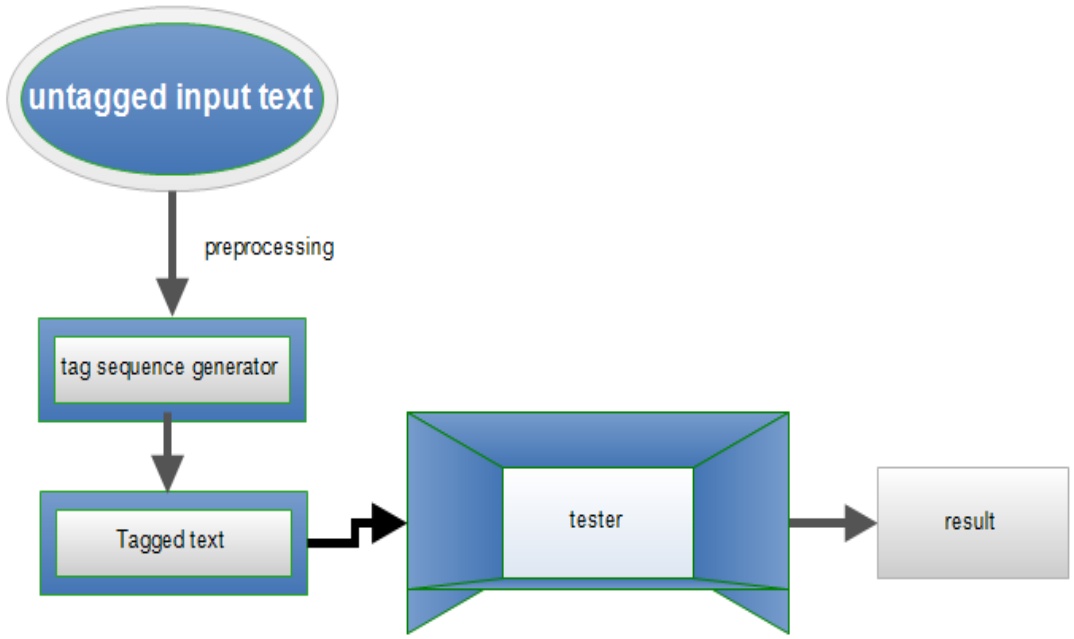


Figure 3.2. HMM Evaluation process and tagger

The untagged Awngi text is given to the sentence splitter module and tokenizer preprocessing components so as to make ready for tagging by the Tag Sequence Generator. Then, the Tag Sequence Generator selects an optimal part of speech tag sequence for the given word sequences and gives the tagged word sequences as an output. The tagged text is given to the tester component for comparison against a manually tagged (so called reference text) and this component gives accuracy of the tagging by counting the number of correctly tagged words.

The optimal sequence of part of speech tags for a given sequence of words in an input sentence to be tagged can be found using the Viterbi algorithm (Milion Meshesha, Getachew Mamo, 2011). The Viterbi algorithm is a dynamic programming algorithm that finds the optimal path in the tagging process. It reduces the complexity of the HMM core issue, finding the best part of speech tag sequence for a given sequence of words in the input sentence, to polynomial time and the algorithm is linear in the number of words to be tagged. (Teklay, 2010) In simple terms, the Viterbi algorithm calculates the probability of all possible paths of the word tag pairs in the input sentence. Afterwards, it will select the path of the word tag pair with the highest probability to be the best path. It uses the lexical and contextual probabilities obtained from the lexical and contextual model to find the best path.

CHAPTER FOUR: RESULT AND DISCUSSION

After we have identified the possible tag sets of awngi language with Experts in the area, Namely Mr. Ayenew W. and Mr. Tesfaye, we did manual tagging for training set. The data set is then divided into two sets: the training set and the testing set. The former one comprises 90% of the corpus while the remaining 10% are used for testing purpose. In this chapter, the detail experiments conducted for this thesis work are discussed briefly.

To evaluate the model performance of the tagger we have used 10 fold cross validation. It will split the training set into 10 folds when K = 10 and we train our model on 9-fold and test it on the last remaining fold. We can make 10 different combinations of 9-folds to train the model and 1-fold to test it. Like this, we can train the model and test them all on 10 combinations of training and test sets.

What we can do afterwards is to take the average of different accuracies up to 10 evaluations and also compute the standard deviation to have a look at the variance.

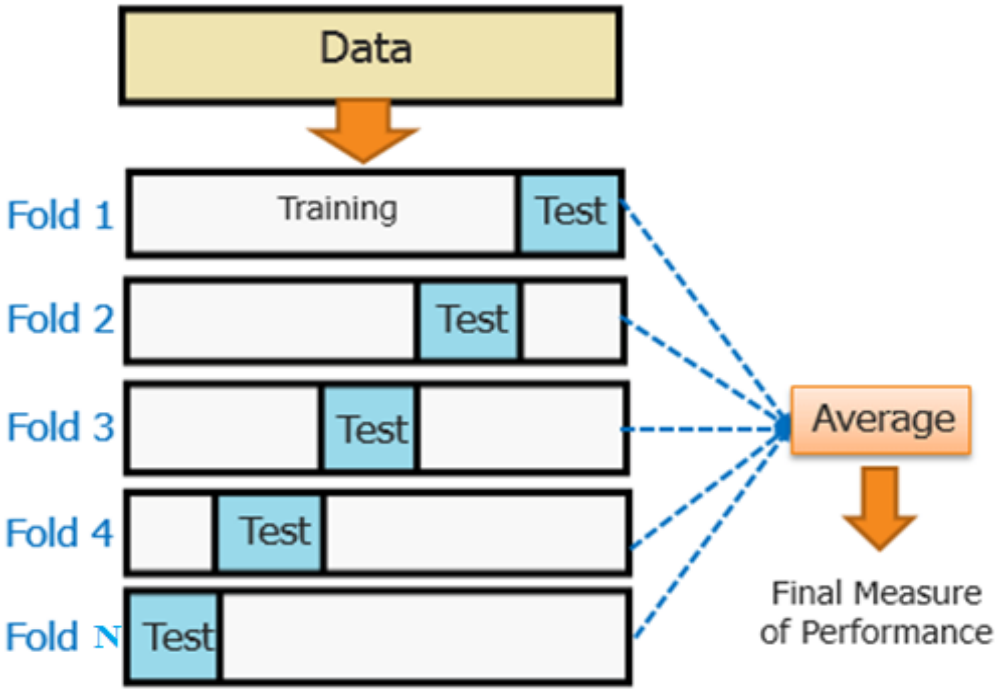


Figure 4.1. fold cross validation

4.1. System performance testing experiments with HMM

The entire training data set was divide into ten equal sizes (each size is 10% of the total training set).The accuracy of the tagger was tested starting by the first 10% of the data and repeating the process by adding 10% to the previous data until the entire training corpus(100%)is used. For every 10 % data added the accuracy variation is recorded.

Average Fold Accuracy = sum(f1+f2+f3+f4+f5+f6+f7+f8+f9+f10)/10

Table 4.1. Shows the accuracy of the system for a given percentage of training data and the error rate occurred in the tagging process.

cross validation fold	Fold Accuracy	Error rate
Fold 1	87	13
Fold 2	91	9
Fold 3	81	19
Fold 4	91	9
Fold 5	88	12
Fold 6	93	7
Fold 7	95	5
Fold 8	95	5
Fold 9	94	6
Fold 10	98	2
Average	91.3	8.7

Table 4.1. 10 fold cross validation performance evaluation and error rate result

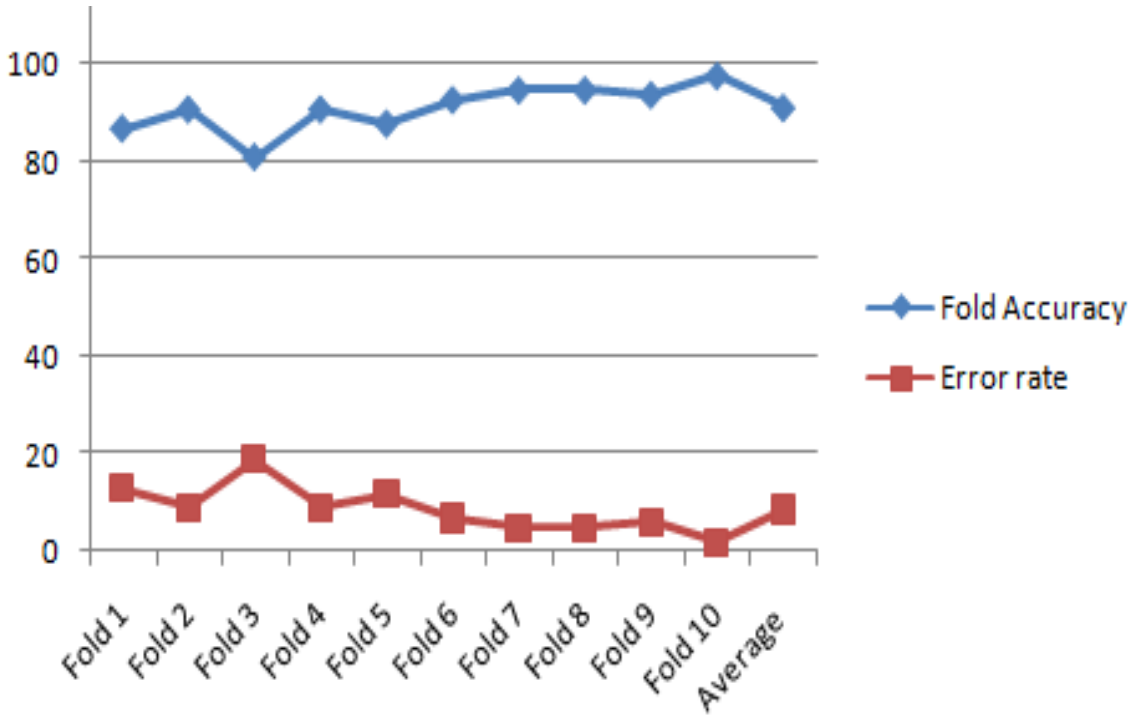


Figure 4.2. 10 fold cross validation and error rate

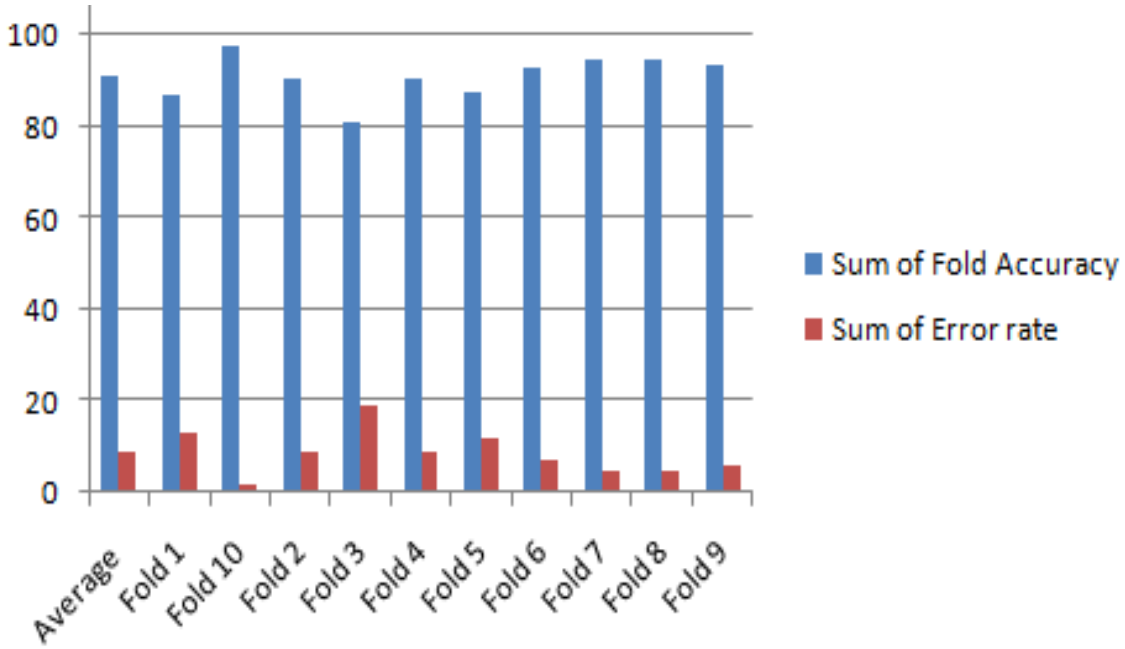
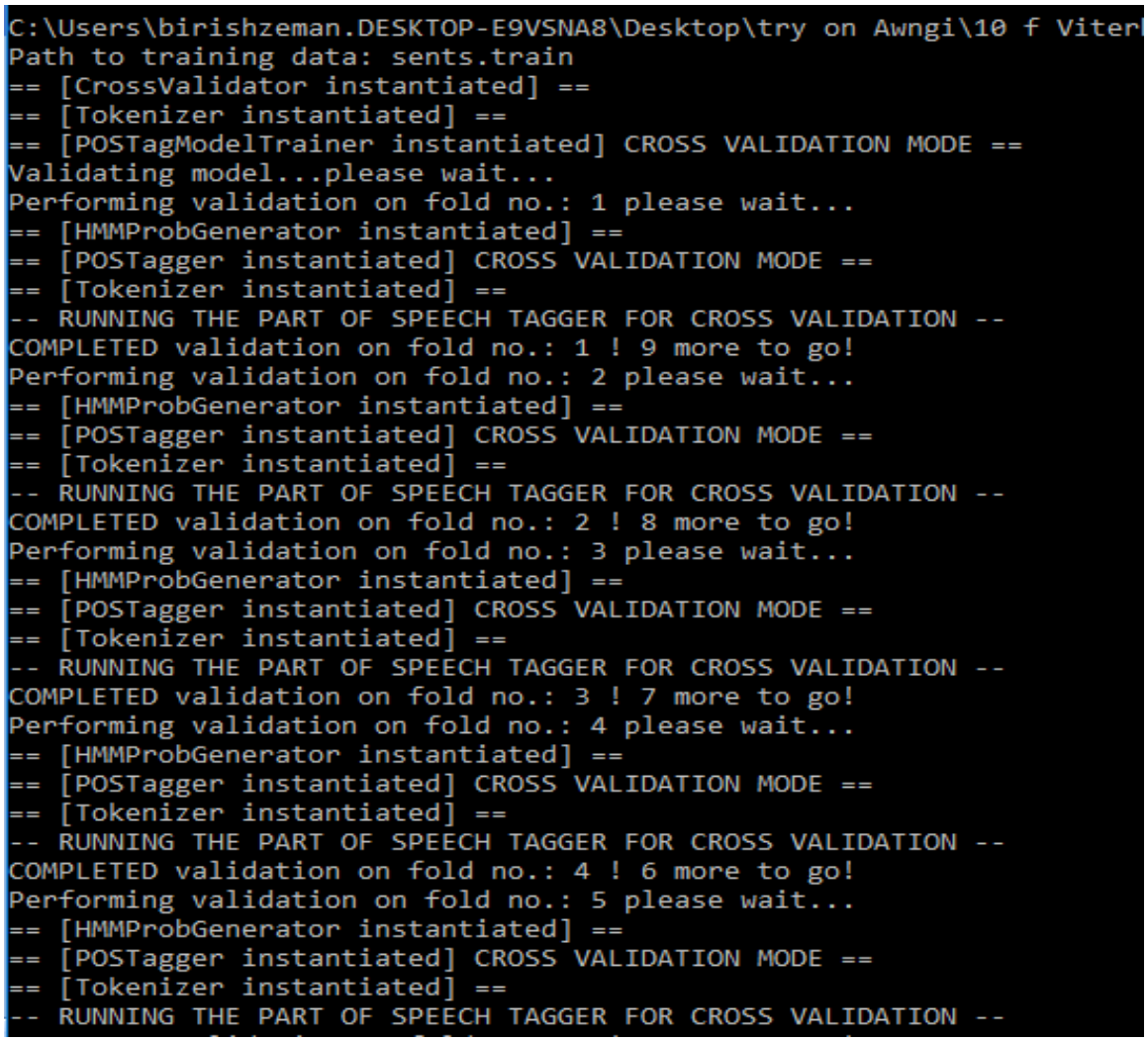


Figure 4.3. 10 fold cross validation and error rate performance evaluation and error rate

The following image describes the average Awngi text tagger performance on python programming language.



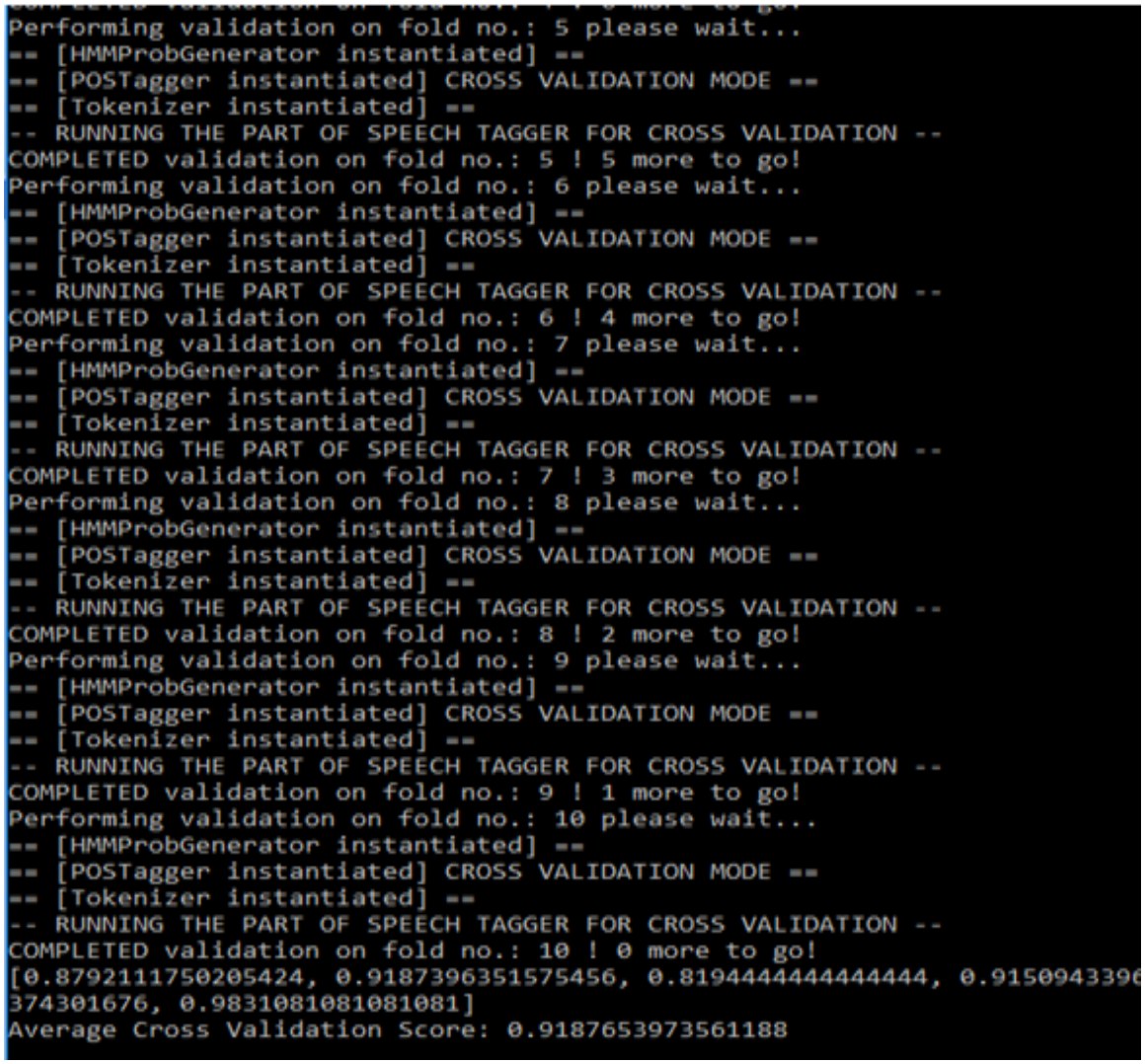


Figure 4.4. Awngi text tagger performance output

CHAPTER FIVE: CONCLUSION, CONTRIBUTION AND RECOMMENDATION

5.1. Conclusion

In this thesis we have briefly discussed about Natural language Processing (NLP) and it's role towards enabling computers to understand natural languages by which most of the human language is recorded. NLP, as it encompasses computational linguistics, important in designing and development of part of speech (POS) taggers, parsers, morphological analyzers etc.

Application areas of part of speech tagging like information extraction, information retrieval; parsing, question answering, speech synthesis, recognition and machine translation were also briefly illustrated. Different approaches used in part of speech tagging like rule based, corpus or stochastic and machine learning (supervised and unsupervised) methods were discussed. The advantages and disadvantages related to each approach were also presented.

In order to understand the approaches used, the language structure and word classes to determine the tag sets, the researcher reviewed different related literatures on the domain area. In addition, related works done for global and local languages particularly on POS tagging were observed.

Based on the literatures reviewed, Awngi is under-resourced language and a lot of Natural language processing tasks are left for researchers.

Among the different types of NLP applications, POS tagging is the primary and the basic research in which the absence of this task on every language will hinder the researchers to conduct high level NLP researches. Due to this the researcher decided to do a research on this area.

In addition to primary and secondary information sources, the research was done by the help of Injibara zone educational administration experts. Starting from understanding the morphological structure of the language, tag set identification and preparing a tagged training dataset, the involvement of different experts on the area was paramount significant to accomplish the research.

In order to do the research we have used a HMM stochastic approach and a total of 24 tag sets identified,450 sentences and more than 3500 tokens(words) have used for training data set. For the tagger performance evaluation purpose we have used tenfold cross validation and the average performance of the tagger becomes 91% which can be considered as significant for the advancement of the language and for the researchers who want to conduct on Awngi and other poor resource local languages.

5.2. Contribution

As it has been discussed on the first chapter, Part of speech tagging is the basic and the primary research that is used as an input for other higher level natural language processing research applications. In this regard, awngi is under resourced local language in which a lot of research areas are open for researchers.

When we come to our research contribution, after a detailed study of Awngi morphological structures, we just collected the necessary data sets which are used for training and testing purpose of the tagger, we have identified the tag sets in order to tag the Awngi sentences. Also the researchers who want to conduct a research on this area this can be used as an input.

5.3. Recommendation

The Awngi Pos tagger which has been investigated and developed is the first attempt for the language and further researches remained and has to be done to improve the performance of the taggers to operational level. In addition, this research has limitations and gaps which can open the door for future researchers to develop POS tagger for Awngi Language that has better performance. Therefore, the following are some of future research directions.

- Due to time and language experts constraint, only tenfold cross validation experiments have been conducted using the stochastic approach. So in the future one can conduct more experiments to improve the performance of the taggers.
- One can do a research with increasing the size of the corpus for training the taggers and increase the performance of Awngi language POS tagger.
- In this research reduction of tags was based on the number of occurrences of tags in the corpus or tags that occur rarely are reduced to the nearest tag but in the

future one can use error analysis to reduce tags and can observe the effect on the performance result of the taggers.

- Since most of the Ethiopian languages are under-resourced and do not have large size POS annotated corpus, one can develop POS Taggers for other local languages following approaches used in this thesis especially on Neural Network approaches, Hybrid Approaches etc.

REFERENCE

Hansan M., Levenberg-M.. (2014). Learning Neural Network ForPart-of-Speech Tagging of Arabic Sentences. Volume 13.

Adafre, S. F. (2005). Part of Speech tagging for Amharic using Conditional Random Fields . Workshop , 47–54.

Ajai Kumar , Shashi Pal Singh. (2014). Hybrid approach for Part of Speech Tagger for Hindi language.

Anderson, S. A. (1988). Morphological theory in Part of Linguistics. Linguistics , 146.

Ankur, P. (2009). Part-Of-Speech Tagging using neural network International Conference on Natural Language Processing.

Anne Kao,Steve Poteet:. (n.d.). Text Mining and Natural Language Processing – Volume 7, Issue 1. Introduction for the Special Issue .

Berhanu, H. (2015). Part of speech tagging for Wolaita language. Addis Ababa: Addis Ababa University.

Bird, S. (2006). the natural language toolkit. In Proceedings of the COLING/ACL on interactive Presentation Sessions. Association for Computational Linguistics. Morristown.

Bjorn Gambackand, Fredrik Olsson, Atelach Alemu Argaw, Lars Asker. Methods for Amharic Part-of-Speech Tagging.

Brill, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: a Case Study in Part-of-Speech Tagging, Department of Computer Science, Association for Computational Linguistics. Johns Hopkins University.

Brill, Eric. (1992). A Simple Rule-Based Part of Speech Tagger In Proceedings of the Third ACL Applied NLP. Philadelphia , 152-155.

Christopher D., Manning Hinrich . (2000). Foundations of Statistical Natural Language Processing. London: MIT press cambridge .

D. Jurafky ,J. H. Martin. (2006). Speech and Language Processing, An introduction to natural Language Processing, Computational Linguistics, and speech recognition. New Jersey.

Daelemans, W., J. Zavrel, and P. Berck. (1996). Part-of-speech tagging for Dutch with MBT, a memory-based tagger generator. dutch .

Daniel J. et al. (2018). An Introduction to Natural Language Processing,Computational Linguistics, and Speech Recognition. stanford: stanford university.

emiru, g. (2016). Development of part of speech tagger using hybrid. addis ababa: addis ababa university .

Fahim Muhammad et al. (2006). Comparison of Different POS Tagging Techniques (n-grams, HMM and Brill‟s Tagger) for Bangla,. conference paper , 4-14.

ganta, b. h. (2015). Part of speech tagging for wolaita language. Addis Ababa: Addis Ababa Univesity.

Getachew, M. (2001). Automatic Part of Speech Tagging for Amharic: An Experiment Using Stochastic Hidden Markov (HMM) Approach. Masters thesis, Addis Ababa University.

Getachew, M. (2009). Part Of Speech Tagger for Afaan Oromo Language. Addis Ababa: Addis Ababa university.

HASAN MUAIDI,Levenberg-Marquardt. (2014). Learning Neural Network ForPart-of-Speech Tagging of Arabic Sentences .

Haykin, S. (1994). Neural Networks: A Comprehensive Foundation, 2nd ed.

Holen, G. L. (2009). HMM tagger for Swidish. Swiden : Swiden .

Hussen, M. (2010). Part Of Speech Tagger for Afaan Oromo Language using Transformational error driven learning (TEL) approach. Addis Ababa: Addis Ababa University.

James, A. (1995). Natural language Understanding. . Technology .

Jedlink. (2019, march 21). gerjanos. Retrieved 2019, from http://jedlik.phy.bme.hu/~gerjanos/HMM/node2.html/

Joswig, Andreas. (2010). "The phonology of Awngi. SIL Internationa.

Levent Altunyurt , Zihni Orhan. (2006). Part of Speech Tagger for Turkish, Masters Thesis, Computer Engineering, Bo_azici University. Bo_azici: Bo_azici University.

Mamo, G. (2009). Part-of-Speech Tagging for Afaan Oromo Language. Masters thesis, Addis Ababa University. Addis Ababa: Addis Ababa University.

Marcos G. et al . (2014). PoS-tagging the Web in Portuguese. National varieties, text. Cilenis Language Technology , 95-101.

Martha Yifiru Tachbelie ,Wolfgang Menzel. (2009). Amharic Part-of-Speech Tagger for Factored Language Modeling. International Conference RANLP , 428–433.

Martha Yifiru, Tachbelie Solomon, Teferra Abate , Laurent Besacier. Part-of-Speech Tagging for Under-Resourced and Morphologically Rich Languages – The Case of Amharic Conference on Human Language Technology for Development. (pp. 2-5). Egypt: Alexandria.

Milion Meshesha, Getachew Mamo. (2011). Parts of Speech Tagging for Afaan Oromo. International Journal of Advanced Computer Science and Applications, , 5.

Mohammed, H. (2010). Part Of Speech Tagger for Afaan Oromo Language using Transformational error driven learning (TEL) approach. Addis Ababa: Addis Ababa University.

Parikh, A. (2009). Part-Of-Speech Tagging using neural network. International Conference on Natural Language Processing.

Peter A. Heeman James F. Allen. (n.d.). incorporating POS tagging in to language modeling. Technopole .

python.org. (2018, november 10). http://www.python.org. Retrieved november 10, 2018, from www.python.org: : http://www.python.org/

Ravi Narayan, S. Chakraverty, V. P. Singh. (march 2014). Neural Network based Parts of Speech Tagger for Hindi, . Third International Conference on Advances in Control and Optimization of Dynamical Systems, (pp. 13-15).

Richard P, Lippmann. (1988). An Introduction to Computing with Neural Nets. IEEE .

Sandipan Dandapat, Sudeshna Sarkar ,Anupam Basu. (2004). A Hybrid Model for Part-of-Speech Tagging and its application to Bengali. Journal of information technology .

Schmid, H. (1994). Part-of-Speech Tagging With Neural Networks Proc. 172-176.

Schmid, Helmut. (1994). Probabilistic Part of Speech Tagger using Decision Trees, Institute of Natural Language Processing. Germeny: university of Stuttgart.

simon, s. (2000). part of speech tagger for swedish. lund university.

SIUM, M. A. (2016). automatic part-of-speech tagger for tigrigna language a hybrid approach. addis ababa : Addis Ababa University .

Solomon. (2008). Automatic Amharic Part-of-Speech Tagging Using Hybrid Approach (Neural Network and Rule-Based). Addis Ababa: Addis Ababa University.

Teklay, G. (2010). part of speech tagger for tigrigna language. Addis ababa: Addis Ababa University.

Tewolde, T. (2002). A modern grammar of Tigrigna. G. Savonarola Roma.

Tsegaye, M. (2013). Developing a Stemming Algorithm for Awngi Text: A longest match approach. Addis Ababa: Addis Ababa University.

Willet, A. N. (2002). Stemming of Amharic words for Information Retrieval. In Literary and Linguistic Computing , 1-17.

Appendices

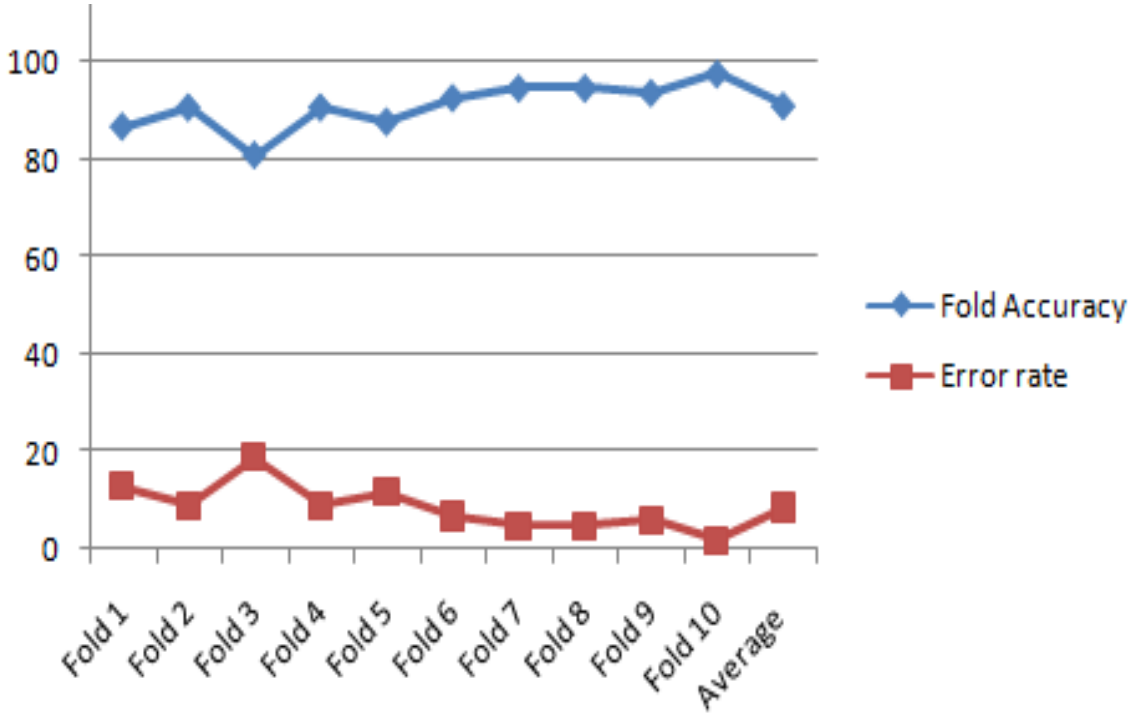
Appendex A: Sample data set(untagged)

ሴላሙ ፎረም እምፕልቶ ያጋማጌኽ ። ክንቲ ውላ ዱቡጊ ቹዋኽ ። ልግስ ታክስስቱስ ክንትሻስ ቤዳራማ/ ታክስሻ ቱስቴ ። አሌሙ ስልታኒው ቹዋ ስር ክንቲ አኽሻኪላ ውላዋጊ አስሜምሻፄ ። ሌጌሱንኩ እናኩ አጌርካው ልግሱ ስር ክንቲ አኽኼ ካቢትፓማ ዙሚስቱስ እንኩዋኽኔ ። ኬራውሳ አግልግሎቴ ኽይፂ እላአኽሻ ቲንዲ ችግሮ ያጋማጊ አኽኼ አዊ ቤረሰብ ቺፅሻፂ ዚጋሚ ወረዲ ክላጅ ኬቴሙዝኩዋንትካ ዙሙና ። አማኻሪ ክልልዲ 3ዎ6 ሻይ የክታር አኹኪ ብቴ ኬሹንኩስ ቴክኖሎጅካ ሊሚፅሻስ ካሊኼ ክልሉ ግብርኒው ቢሩ ጌሌፅኾ ። ሳውዱያ አረቢያ ትክላይ ሚኒስትር ድክተር አብይ አህመዴስስታ ፕሬዝዲንት ኢሳያስ አፈወርኪስ ንጉሳዊ ክብሩሳ ሜዲልያዋ ሼሌምትኾ ። እላፌእጉሳ ፂፄ ፌፄሙንኩ ታራታርስታንትካዋ ፃንኩፅካማ እግስ ዱጋንታስታ ፋይፃንቲ ክፃት ፄውስታት ኪላ ጌሌፅካ ። ፄትቲኩ አካልካ ኪላ ጉዲዮ እምቢትንካማ ፃንኩፃንታስታ ፋይፃውሳ እጋዊ ክሳቶ ካፃንታኪላ ጉሴ እይካ ። ቻይና ካቢትኚስ አሜሪካስ ዴርብስታንኩ አይኪ አይኔት ግብሮ ኬዌምሻታት አኽሻ ጌሊፂስ እጂቴ ።

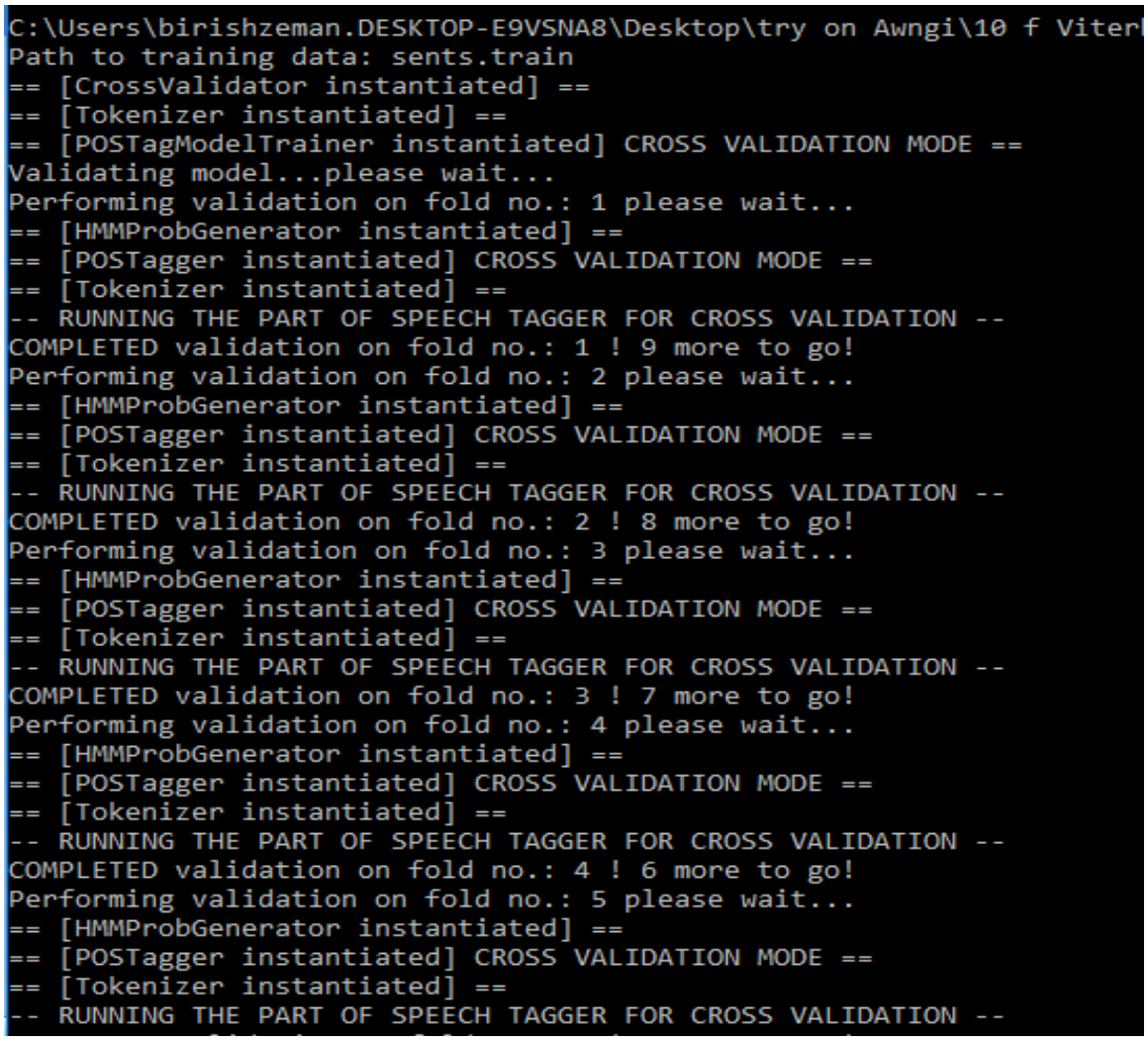
Appendix B: Training set (tagged)

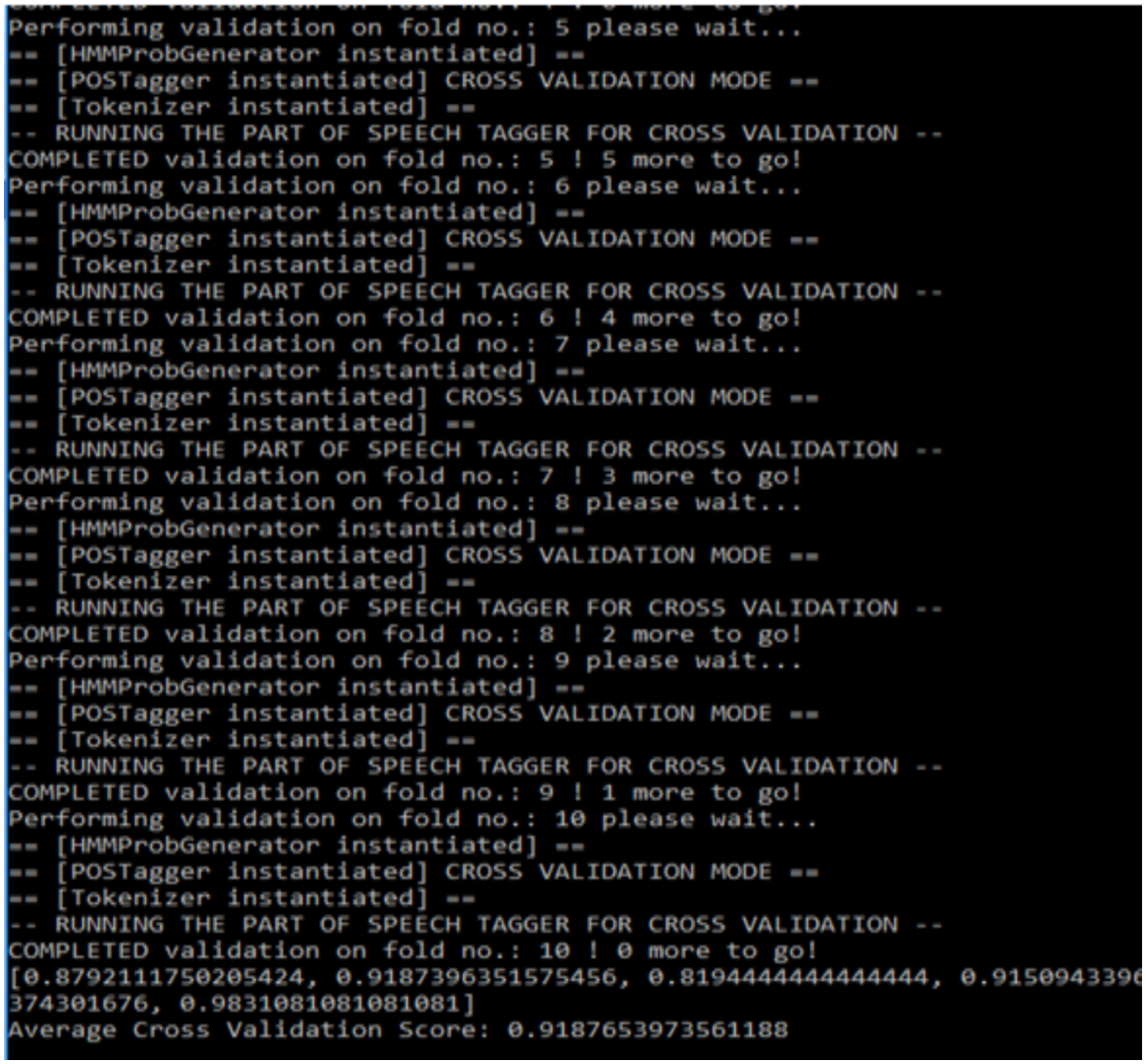
ሴላሙ/nn ፎረም/nn እምፕልቶ/nn ያጋማጌኽ/v ።/punct ክንቲ/nn ውላ/adj ዱቡጊ/nn ቹዋኽ/v ።/punct ልግስ/nn ታክስስቱስ/vv ክንትሻስ/nn ቤዳራማ/vv ታክስሻ/vv ቱስቴ/vv ።/punct አሌሙ/nn ስልታኒው/adj ቹዋ/Adj ስር/adj ክንቲ/nn አኽሻኪላ/conj ውላዋጊ/adj አስሚምሻፄ/vv ።/punct ሌጌሱንኩ/vv እናኩ/vv አጌርካው/nn ልግሱ/Adj ስር/adj ክንቲ/nn አኽኼ/conj ካቢትጋማ/adv ዙሚስቱስ/vv እንኩዋኽኔ/vv ።/punct ኬራውሳ/nn አግልግሎቴ/nn ኽይፂ/nn እላአኽሻ/vv ቲንዲ/nnprep ችግሮ/nn ያጋማጊ/vv አኽኼ/conj አዊ/nn ቤረሰብ/adj ቺፅሻፂ/nn ዚጋሚ/nn ወረዲ/nn ክላጅ/nn ኬቴሙዝኩዋንትካ/nnadj ዙሙና/vv ።/punct አማኻሪ/nn ክልልዲ/nn 3ዏ6/cd ሻይ/nn የክታር/nn አኹኪ/conj ብቴ/nn ኬሹንኩሰ/Adj ቴክኖሎጅካ/nn ሊሚፅሻስ/vv ካሊኼ/vv ክልሉ/nn ግብርኒው/nn ቢሩ/nn ጌሌፅኾ/nn ።/punct ሳውዱያ/nn አረቢያ/nn ትክላይ/adj ሚኒስትር/nn ድክተር/adj አብይ/nn አሀመዴስስታ/nnprepconj ፕሬዘዲንት/adj ኢሳያስ/nn አፈወርኪስ/nnprep ንጉሳዊ/adj ክብሩሳ/adjprep ሜዲልያዋ/nn ሼሌምትኾ/vv ።/punct እላፈእጉሳ/adj ፂፄ/nn ፈፄሙንኩ/vv ታራታርስታንትካዋ/nn ፃንኩፅካማ/vv እግስ/nnprep ዱጋንታስታ/vvconj ፋይፃንቲ/adj ክፃት/nn ፄውስታት/adj ኪላ/prep ጌሌፅካ/vv ።/punct ፄትቲኩ/nnprep አካልካ/nn ኪላ/prep ጉዲዮ/nn እምቢትንካማ/adv ፃንኩፃንታስታ/vv ፋይፃውሳ/nn እጋዊ/adj ክሳቶ/nn ካፃንታኪላ/vvprep ጉሴ/adj እይካ/vv ።/punct ቻይና/nn ካቢትኺስ/adv አሜሪካስ/nnprep ዴርብስታንኩ/vv አይኪ/adj አይኔት/adj ግብሮ/nn ኬዌምሻታት/nn አኽሻ/vv ጌሊፂስ/adj እጂቴ/vv ።/punct

Appendix C: 10 fold cross validation performance evaluation and error rate



Appendix D: Average Result of tagger Accuracy




Performing validation on fold no.: 5 please wait...
== [HMMProbGenerator instantiated] ==
== [POSTagger instantiated] CROSS VALIDATION MODE ==
== [Tokenizer instantiated] ==
-- RUNNING THE PART OF SPEECH TAGGER FOR CROSS VALIDATION --
COMPLETED validation on fold no.: 5 ! 5 more to go!
Performing validation on fold no.: 6 please wait...
== [HMMProbGenerator instantiated] ==
== [POSTagger instantiated] CROSS VALIDATION MODE ==
== [Tokenizer instantiated] ==
-- RUNNING THE PART OF SPEECH TAGGER FOR CROSS VALIDATION --
COMPLETED validation on fold no.: 6 ! 4 more to go!
Performing validation on fold no.: 7 please wait...
== [HMMProbGenerator instantiated] ==
== [POSTagger instantiated] CROSS VALIDATION MODE ==
== [Tokenizer instantiated] ==
-- RUNNING THE PART OF SPEECH TAGGER FOR CROSS VALIDATION --
COMPLETED validation on fold no.: 7 ! 3 more to go!
Performing validation on fold no.: 8 please wait...
== [HMMProbGenerator instantiated] ==
== [POSTagger instantiated] CROSS VALIDATION MODE ==
== [Tokenizer instantiated] ==
-- RUNNING THE PART OF SPEECH TAGGER FOR CROSS VALIDATION --
COMPLETED validation on fold no.: 8 ! 2 more to go!
Performing validation on fold no.: 9 please wait...
== [HMMProbGenerator instantiated] ==
== [POSTagger instantiated] CROSS VALIDATION MODE ==
== [Tokenizer instantiated] ==
-- RUNNING THE PART OF SPEECH TAGGER FOR CROSS VALIDATION --
COMPLETED validation on fold no.: 9 ! 1 more to go!
Performing validation on fold no.: 10 please wait...
== [HMMProbGenerator instantiated] ==
== [POSTagger instantiated] CROSS VALIDATION MODE ==
== [Tokenizer instantiated] ==
-- RUNNING THE PART OF SPEECH TAGGER FOR CROSS VALIDATION --
COMPLETED validation on fold no.: 10 ! 0 more to go!
[0.8792111750205424, 0.9187396351575456, 0.8194444444444444, 0.915094339
374301676, 0.9831081081081081]
Average Cross Validation Score: 0.9187653973561188