Amharic Phrase Chunking with Conditional Random Fields
Alemebante Mulu\textsuperscript{1}, Vishal Goyal\textsuperscript{1}

\textsuperscript{1}Department of Computer Science, Punjabi University Patiala, India. alembantemulu184@gmail.com, vishal.pup@gmail.com

Abstract: This paper particularly presents a Conditional Random Field (CRF) based phrase chunking system for Amharic language. Chunking is the series of actions which has divided or segmented the sentence into phrase by the arrangements of correlated word groups. Therefore, chunking is recognized by the identity of chunk labels and the boundary that describes chunks. In this research work, our goal is to develop chunking for Amharic language by using different tagging schemes for identifying the chunk boundaries and incorporate the tags in form of contextual information. Therefore we have identified the problem, common setting, solution and improvement of chunking as well. In addition to this, we have also constituted the special word, Part-of-Speech (POS) tagging information, morphological analysis as an input to increase the performance of the system. Totally, we are using 400,000 tagged words and have evaluated the result by the combination boundary identification and labeling. Since we are using large amount of tagged data for Part of Speech tagging, it although performs good results with chunking. Finally, the average accuracy of the system is reached to 94.2\%.

Keywords: Chunk, CRF, HMM, NLP, POS.

I. INTRODUCTION

A. Background

Natural language processing (NLP) is one large area or sub field of computing science that can be written, spoken, read and listened by human beings for sending or receiving information via computer devices (Lise & Ben, 2007). The pleasing communication between human beings and computer machine is performed with the help of Natural Language Processing (NLP) technology and its application to the human-computer interaction is most easy (Jackson & Moulinier, 2007). Natural Language processing (NLP) understands the requirement and the general language structure of text at phonological, at morphological, at syntactical, at semantical, at discourse and pragmatic levels (Jurafsky & H.Martin, 2019) to make, increase the instance and performance of NLP applications that have been done at different levels.

Having clear opinions, the combination of two or more phrases in a sentence is very useful to construct the full sentences and phrases by themselves for a variety of Natural Language Processing (NLP) application, and the process which directly labels the small group of words is called phrase chunking. Grover &Tobin (2006) states that considers chunking is able to be used a practical for Named Entity Recognition attentively.

Chunking is one of Natural Language Processing (NLP) task that have been used to allocate phrases by their label group of contiguous words. We can say that it’s important input for constructing a sentence as well as identifying the grammatical classes. Whereas, Part of Speech (POS) is used as an input and it needs to help in developing the chunking system, especially to get a better accuracy. In addition to this, the task of it helps to make it easier for improving the efficiency of the subsequent processing like parsing and grammar chucking. Part of Speech tagging (POS) and phrase chunking are carried out almost with the same piece of work that performs, the only difference is part of speech tagging which has as an essential feature of word categorization and chunking is the place in a particular phrase level category.

Chunking is very useful for identifying the phrase level under the Natural Language Processing (NLP) application for a diverse of language, and the task labels the phrase to categorize under the system. Therefore, chunking application can consist in dividing the sentence in to phrase level in syntactically correlated parts of the word (Ibrahim & Assabie, 2013).“The syntactic level deals with analyzing a sentence that generally consists of segmenting a sentence into words and recognizing syntactic elements and their relationships within a structure” (Ibrahim & Assabie, 2014, p.297).

Phrase chunking is the stream order that helps to identify and analyze well organized data. In addition to this it is used for further investigation under the natural language processing application development like grammar checker, information extraction, information retrieval, name entity recognition and
other related tasks decidedly, chunking has begun to be an interesting alternative of one or more things to full parsing. The arrangement of the sentence is categorized on the habit the linguistic structure or phrase chunking and its one big task is natural language processing. Hence, we describe the above statement that aim of phrase chunking is to divide a sentence in to certain syntactic units phrase level. For example, the sentence “መፍቅስቂት ከሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ በግራ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የበትልቅ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የበትልቅ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የበትልቅ የሆነውን የ alanı የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የቡድ በግራ የሆነውን የantu የሆነውን የቡድ በግራ የሆነውን የቡድ የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የantu የሆነውን የardu የሆነውን የardu የሆነውን የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የardu የaru
approaches (Ali & Hussain, 2010; Xu, Zong, & Zhao, 2006) like Conditional Random Field (CRF), Support Vector Machine (SVM) and Hidden Markov model (HMM) (Pranjal, Delip & Balaranam, 2006). On the other hand different methods have also been investigated particularly on boundary identification and chunk labeling. Most of the earliest chunking systems have used standard HMM based tagging methods in modeling the chunking process. English chunking system has been done by (Church K, 1988) and it is used by HMM tagging methods with statistical approach. In this work, I have used the special word and part of speech and tagged text as used as an input. In addition to this we implement by using the CRF based Chunking mechanism. For further clarification this research work has focused on identification of the tag set types. Inside, or Outside of the given chunk has become the de-facto standard for this task (Yoav, Michael & Meni, 2006). The combination of words, special word and POS tags that gives the best result for this research work. Also, different methods are used and compared in order to get best results for labeling the chunks. The observation of fact in practical contact can be used to develop Chunkers for other languages as well.

II. CONDITIONAL RANDOM FIELD (CRF)

The degree of recognizability in outline of a Conditional Random Field (CRF) is CRFs based on the idea of Markov Random Fields or the extension of HMM and Maximum-Entropy Models. Conditional Random Field (CRF) is a form of recognized modelling that has been successfully used in different spheres such as part of speech tagging and other Natural Language Processing tasks. “Point out that each of the random variable label sequences Y conditioned on the random observation sequence X” (Xu, Zong & Zhao, 2006, pp. 87293). Generally the bottom-up process of conditional random field is applicable on the combination of multiple features of the data and it can be putting together the probability of sequential order in the given data can be represented as P (Sequence | Data). Therefore the conditional random field is to be able to represent a graph G = (V, E), but not consider a chain. Then, it’s conditioned on X, observation sequence variable and each node represents a value Yv of Y output label. Let G be a factor graph of Y, then P (Y|X) is a conditional random field where y is a label, and x an observation sequence (Avinesh, 2007; Himanshu & Anirudh, 2006). Based on the above concept the premises of Hammersley-Clifford theorem states that a random field is an MRF if it can be described in the form below. The exponential is the sum of the clique potentials of the undirected graph

\[ P(Y|X) = \frac{\exp \sum_c (\lambda_i l_i(x, y_r) + \sum_j \mu_j g_j(x, y_r, y_{r-1}))}{Z(x)} \]

For further explanation we try to give a detail of each symbol as appointed to member of the above formula. Therefore, the symbol “c” is represented by the State Feature Weight \( \lambda = 10 \) and its one possible weight value for this state feature is strong, the symbol “f” is represented by the State Feature Function \( f(x \text{ is stop}), f/ ) \) and its one possible state feature function for our attributes labels, the symbol “\( \mu \)” is represented the Transition Feature Weight \( \mu = 4 \) and its one possible weight value for this transition feature and the final symbol “\( \tau \)” can be represented the transition feature function \( g(x, /iy/, /k/) \) and its one possible transition feature function indicates /k/ followed by /iy/.

A. Approach

Most of the time Natural Language Processing (NLP) application has been connected by the flexible series chain. Along with the combination of application can form by putting parts together. When the developer train to the Part of Speech (POS) tagger application for Amharic language it may use the Amharic morphological analyzer, tokenization and properly tagged taring corpus to get the root-word and possible POS tags for every word in the corpus. By using tokenization they can by constricting the root-word, morphological analyzer and suggest other POS tags information like prefix, infix, suffixes, word length indicator, sentence boundary, punctuation mark identifier and presence of special characters which is added to the training data. The POS tag assigned to every token is used to discover these positions as well it could be the most desirable input to develop excellent chunking application for every language. Before we implement the system for chunking system, it needs to manage the system on unannotated input which, a similar data (corpus) and has to be prepared before the system can predict. To recognize the chunks, it is required to be found the place where a chunk can start the new chunk and ending of the chunk can be completed (Himanshu & Anirudh, 2006). In addition to this, the system is used for two training phase approaches. The chunking set of rules is to be followed into two phases, namely: Chunk boundary identification and Chunk label identifier. We first extract chunk boundary recognizer and chunk label recognizer for each word in the corpus. In the first phase chunk tags (Chunk Boundary Recognizer-Chunk Label Recognizer) are designated to each single distinct meaningful element in the training data and the data is trained to predict the corresponding CB-CL tag. In the second phase, we teach a particular skill for the system on the above feature template to make known beforehand the chunk boundary recognizer (CB). Finally chunk label recognizer (L) from the first phase and the chunk boundary recognizer from the second phase are combined together to obtain the chunk tag. To identify the Chunk Boundary Recognizer (CB) by itself we classify it in to four main parts, namely:

- BGN (Beginning the chunk word),
- ISD (Inside the chunk word),
- STP (Stop the chunk word),
- BGN-STP (Beginning-Stop the chunk word) and
- OSD (Outside the chunk word)

The second chunk label recognizer is Chunk Label Recognizer (CL) and it can be divided into five parts, namely:

- NP (Noun Phrase),
- VP (Verb Phrase),
- AdjP (Adjective Phrase),
- AdvP (Adverb Phrase) and
- PP (Prepositional Phrase)
Easy to perceive putting chunk label identifier in specifying, a chunk that has been a part of markers at pre terminals.

III. EXPERIMENT

We make an effort to achieve the chunk tags using contextual information. This research work is being used by morphological analysis, special word, POS tagger and we implement by using two phases, namely: Chunk boundary identification and Chunk labeling.

The starting point of this research work is carried out by the combination of morphological analysis, POS tagger and chunking tag using CRF of the chunk tag schemes are discussed in the beginning section. Therefore we use the arrangement of chunk tags considered by special word, POS-Tag: Chunk-Tag, which are to make adequate by capable of successfully reaching the intended target.

A. Part of Speech (POS) Tagger

First of all POS tagger is the greatest significance task to develop chunking and it needs the train corpus either for developing POS tagger system or the chunking system. Therefore the POS tagging corpus should be trained with a basic template. For this reason the language expert doing process manually tagged data for corpus preparation as well. The second step for developing POS tagger is to recognize the error and try to handle the base problem like morphological analysis for recognizing the root-word. By nature Amharic language is morphologically rich and sometimes it has attached the right way to the word, it was felt to be used for prefix or suffix information. Prefix is put together for every word as the first two or three characters and suffixes were put together for every word as the last two or three characters of the word. When we compare the Amharic language to the other one like English language, the words that are categorized under proper nouns are used for capitalization and that needs to help to recognize them. However, there is no such mark is done in Amharic language and we watch attentively such kind of problem can be solved by using very modern application like Amharic Horn-Morphism.

For being associated with this research work we used the POS tag set that has been developed by using “The Annotation of Amharic News Documents” project at the Ethiopian Language Research Center. The determination of the project was to manually tag each Amharic word in its context (Abney, 1992). A new POS tag set for Amharic has been obtained from this project. In this project, the basic tag set are listed below. The tag set has 11 basic classes and 63 derived tag sets: Nouns (N), Verbs (V), prepositions (PREP), adjectives (ADJ), adverbs (ADV), pronouns (PRON), conjunction (CONJ), interjection (INT), punctuation (PUNC), numeral (NUM) and UNC which are in a particular position for unclassified words and used for words which are difficult to place in any of the classes. Some of these basic classes are further subdivided and a total of 63 POS tags have been identified.

B. Chunking

“Chunking is the task of identifying and segmenting the text into syntactically correlated word groups” Dhanalakshmi, Padmavathy, Anand, K.M., Soman, & S, 2009, pp. 436-438).

In this research work we apply two basic processes and these are used to help accomplish the chunking system. First, the POS annotated corpus is made ready for use and the second step is prepared as basic arrangement or sequence of process (Chandan, Vishal & Umrinderpal, 2015). Those processes are the Chunk Boundaries (CB), Chunk Label (CL) and both Chunk Boundary-Chunk Label (CB-CL) chunk tags. Let’s say, Part of Speech (POS) tags is represented by the tag set “T” and the chunked data is represented by “C”, then \( T_n = (t_1, t_2, \ldots, t_n) \) where \( T \) is a sequence of \( t_i \) from POS tag set and \( C_n = (c_1, c_2, \ldots, c_n) \), \( c_i \in C \) where \( C \) is a sequence of \( c_i \) to \( c_n \) chunk tags. Therefore, the problem can be solved by the cooperation of chunk tag sequence (C) and the sequence of POS tag sequence (T). Since we are using the Conditional Random Field (CRF) approach, the probabilistic model is solve corresponding sequence of Part of Speech (POS).

The best for identification of Chunk Boundary is combination of <POS word_POS><chunk_tag>→<POS_tag> and the subsequent conversion to 2-tag set gives better results (Ashish, Arnab & Sudeshna, 2005).

Since we stand to solve the problem we try to identify the basic skeleton of the solution and it should be the sequentially. The basic skeleton of this research work is Problem, common setting, solution, improvement and the output of the chunking system

\[ \text{Problem} \rightarrow \text{Small vocabulary} \]
\[ \text{Common Setting} \rightarrow \text{Input sentence: words} \]
\[ \text{Input/output tags: POS} \]
\[ \text{Solution} \rightarrow \text{Input sentence: POS} \]
\[ \text{Input/output tags: POS + Chunks} \]
\[ \text{Improvement} \rightarrow \text{Input sentence: Special words + POS} \]
\[ \text{Input/output tags: Special words + POS + Chunks} \]
\[ \text{Chunking} \rightarrow \text{Input sentence: POS} \]
\[ \text{Input/output tags: Chunks} \]

\[ f_c(w_i, p_i, c_i) = \begin{cases} (w_i, p_i, w_i, c_i) & w_i \in W_s \\ (p_i, c_i) & w_i \notin W_s \end{cases} \]

In practice: put the output of the chunking system Modification of the Input data (ENA train and test).

Finally, our aim is to calculate the probability of POS tagging \( (T_n) \), chunking \( (C_n) \) and word \( (W_n) \). Most probable chunks of the sequence \( W_n \). The chunks are marked with chunk tag sequence \( C_n = (c_1, c_2, \ldots, c_n) \) where \( c_i \) stands for the chunk tag corresponding to each word \( w_i \), \( c_i \in C \).

1) Chunk Boundary Identification

Basically, this research work is to provide for consideration by extracting the chunk boundary identification and chunk Label
markers for each word in the annotated corpus. We can classify the chunk tag and chunk boundary identification in different categories. Phrase chunking can be categorized as follows, like Noun Phrases (NP), Verb Phrases (VP), Adjectival Phrases (AdjP) and Prepositional Phrase (PP) (Yimam, 2009). In spite of that, there are seven boundaries in order to recognize the boundaries of phrase chunk in the given sentences. These are BGN, ISD, STP, BGN-ISP, OSD, < and >. The first formats BGN are complete chunk representation which can identify the beginning phrases. The second formats ISD are complete chunk representation which can identify the inside phrases. The third formats STP are complete chunk representation which can identify the stop phrases. The fourth formats BGN-ISP are the chunk representation which can identify the combination of beginning and stop phrases. The fifth formats OSD are complete chunk representation which can identify the outside phrases. The last format is < and > which represent the initial and final chunk words respectively.

Suppose, we have a sentence as follow: ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንדם ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የእምራት ነው። ይህ ከምስታቸውን የአንድ እምሮ ይህ የethical.
CONCLUSION

Text chunking is an isolated part of information which adds more structure to the sentence used and also referred as part of speech tagging and shallow parsing. Therefore Chunker divides a solid piece of a sentence into phrase by using a label to each chunk. POS tag and chunk in a sentence are absolutely necessary and in great demand for the capability of a computer to automate for further investigation in many approaches in the area of NLP. In this era, the area of NLP application has been conducted for different languages for different purposes such as Question Answering, Morphological Analysis, Text Prediction, Part of Speech tagging, Named Entity Recognition, Parser, Chunking, Clause Boundary Identification, Information Extraction, Grammar Checker for English and other related languages. In addition to this, it is used in different area like linguistic acquisition, Psychology, Sequence Learning, and Memory Architecture, etc.

In this research work, we have the target of acquiring Conditional Random Field (CRF) based Amharic phrase chunking. Our justification behind pick out as being CRF for developing the chunker instead of other models such as HMM, SVM or Maximum Entropy Model. CRF works productively with minimum data proves to be efficient without any specific modification for other languages as well. We dig out several frame works to achieve the objective of this research work as well. For example, Amharic POS tagging, identify the boundary of the phrase, categorized the label of the chunk as the basic build task to achieve this research work.

In this paper, the chunking tagging scheme is categorized in to five boundary identifications it is to the seen if there are five boundary identification, BGN, ISD, STP, BGN-STP and OSD.

The constructed corpus for this research work is tagged manually by the language expert and the rules are generated from Amharic language to support tagged corpus like a supporter tools for this research work as well. Finally, this chunk tagged data is used as an input by the chunker next to the special words and POS tagging. This research work is also formally introduced to the model Conditional Random Field (CRF) and the approach is used to capable the intended target of Amharic text chunking. Basically the CRF model is used to get best results rather than HMM or Maximum Entropy Model. In addition to this the experiments carry out by Python 3.7. System evaluation of the text chunker performance was made based on the evaluation procedures outlined in the thesis. In these paper only one parameter, the percentage of correctly chunker sentences in the sampled text has been used to measure the performance of the chunker. Based on the output, it shows the results are achieved to around 94.2%.

ACKNOWLEDGEMENT

First of all, it is proud his loge to pay obeisance and thank the Almighty God for the completion of this research work successfully. All the time, God made strong enough and this is me everything, the valuable gift from him and it’s really beyond my expectation. Big response, big thanks to all who supported me and all that is up until the present to come.

Next, I have to convey my thanks to the Ministry Education of Ethiopian government who have supported me by finance and covering all educational expenses. Also, I would like to thank my brother and friend, Mr. Sintayehu Hirpassa. Starting from the scratch till to the completion of this work, he was very cooperative and supportive all the time I feel obliged to stay in Punjabi University Patiala and get valuable suggestions throughout this work.

Finally, I would like to thank my family especially for my mom Zoma Mengistu. She has been lending many people and I fell grateful them. Have always supported me during my research work and I am very pleasing for that. Convey my sincere and thankful gratitude to my adviser and other faculty for extending enriched guidance and advice in accomplishing this gigantic task.

REFERENCES


Acknowledgement:

Mrs. Zuma Mengistu, my mother, has supported me and encouraged throughout this work.


***