ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 1, pp: (12-23), Month: January - March 2020, Available at: www.researchpublish.com

# Ge'ez POS Tagger Using Hybrid Approach

HAGOS GEBREMEDHIN GEBREMESKEL

Nankai University, college of Software Engineering, December 2019

*Email: hagosgebrem@gmail.com or hagosgebrem@yahoo.com*

***Abstract:*** **This paper proposes a series of carefully designed a Ge'ez POS tagging using Hybrid approach. Trigram N tag tagger combined with the human written rule, regular expression and morphological pattern analysis based tagger of Ge'ez part of speech tagger. Ge'ez literature on syntax, morphology and grammar are reviewed to understand the nature of the language and also to identify possible tag sets. Experiments aiming at evaluating the influence of automatic pre-annotated on the manual part-of-speech annotation of a corpus, both from the efficiency and the accuracy points of view, with a specific attention drawn to biases. As a result, 26 broad tag sets were identified and 15,154 words from around 1,305 sentences collected from one genre i.e., Holy bible. Then, those words ware manually tagged by Ge'ez language professionals for training and testing purpose.**
**The hybrid of TnT with human annotated rule, regex and morphological pattern analysis of Ge'ez language is assumed to perform better than the TnT taggers taken alone. Individual and hybrid experiments have conducted for the three types of taggers namely the TnT tagger, TnT with Regex tagger and Hybrid tagger. The results are 77.87%, 82.23% and 94.32% performances are obtained for TnT tagger, TnT with Regex tagger and Hybrid taggers respectively. Therefore, the performance of Hybrid approach have the best than individual performance. Finally, this paper concludes Hybrid approach have permissive result for Semitic languages.**

***Keywords:*** **Ge'ez, POS tagger, NLP, TnT, Regex, Hybrid POS tagger.**

## 1. INTRODUCTION

Language is one of the fundamental features of human behavior and it constitutes a crucial component of our lives. In its written form, it serves as a means of recording information and knowledge on a long term-basis and transmitting what it records from one generation to the next. In its spoken form, it serves as a means of coordinating our day-to-day life with others (Allen_James, 1995).

According to Noam Chomsky (Anon., n.d.), a language is a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements. Language is an aspect of human behavior. In written form, it is a long-term record of knowledge from one generation to the next while in the spoken form it is a means of communication. Language is the key aspect of human intelligence and can be categorized as natural and Artificial language. Natural language is an ordinary language that has evolved as a normal means of communication among people. Examples: English, Ge'ez, Amharic, Afaan-Oromo and Tigrigna.

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve: natural language understanding, enabling computers to derive meaning from human or natural language input; and others involve generation 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 in a computer (Liddy_Elizabeth, 2001). Additionally, NLP is the means for accomplishing different types of tasks and/or applications. Such tasks include part of speech (POS) tagging, named entity recognition (NER), information retrieval (IR), speech recognition, machine translation, question answering etc. (Liddy_Elizabeth, 2001).

POS tagging is the process of assigning parts of speech like noun, verb, preposition, pronoun, adverb, adjective or other lexical class markers to each word in a sentence or literature. POS tagging is the first step to understanding a natural language. Most other tasks and applications heavily depend on it (Binyam_Gebrekidan, 2009). The significance of POS

(also known as word classes, morphological classes, or lexical tags) for language processing is that it gives a large amount of information about a word and its neighbor. POS tagging is considered as one of the necessary tools. The accuracy of many NLP applications depends on the efficiency of the POS tagger (Sisay_Fissaha, 2005). POS tagging can be used in text to speech (TTS), IR, shallow parsing, information extraction (IE), linguistic research for corpora (Jurafky & Martin, 2009) and also as an intermediate step for higher level NLP tasks such as parsing, semantic analysis, machine translation, and many more (Jurafky & Martin, 2009) . POS tagging, thus, is a necessary application for advanced NLP applications in Ge'ez or any other languages.

Much of the research in natural language processing has been dedicated to resource rich languages like English, French and other major European and Asian languages. African languages have, however, received far too little attention. In fact, most are being spoken by less and less people. Nowadays Part of Speech tagger is developed for different languages and it remains an intensive research area for other different languages. Among the languages with POS tagger developed are Tigrigna (Yemane_Keleta, et al., July 2016), (Teklay_Gebregzabiher, November, 2010), Amharic (Binyam_Gebrekidan, 2009), Kafi-Noonoo (Zelalem_Mekuria & Yaregal_Assabie, 2013), Arabic (Hadni_Meryeme, et al., December 2013), Afaan-Oromo (Getachew_Mamo & Million_Meshesha, 2011), etc. As to the best of the researcher's knowledge, Ge'ez is then a Language that does not have POS tagger developed so far.

Ge'ez is the classical language of Ethiopia within the Semitic language family. It is grouped under north Ethiopian Semitic along with Təgrä and Təgrinya (Leslau & Wolf, 1987). Ge'ez or Ethiopic was the spoken language until the end of the Axum Empire in the ninth century (Marvin_Lionel_Bender, 1976). Today this language is used only for religious writings and liturgical services in the Ethiopian Orthodox Tewahido Church, Eritrean Orthodox Tewahido Church, Ethiopian Catholic Church and the Beta Israel Jewish community.

## 2. MOTIVATION

Ge'ez is the language of many Ethiopian literatures and manuscripts. Several ancient manuscripts, arts, scriptures, heritages, historical, ethical and religious chronicles that can be used as a primary source of knowledge are found in Ge'ez language (MahibreKidusanResearchCenter, 2010). The ancient philosophy, tradition, history, knowledge etc. of Ethiopia was being written in Ge'ez and also there are different books which are written by this language. These resources can be used as source of philosophy, creativity, knowledge and civilization both to Ethiopia and the rest of the world. To use, keep these resources, and transfer these identities to the next generation, the citizens must understand semantically and syntactically way of these written books/documents. If they didn't know the idea in the documents, they will not give any attention for these heritages. If someone who is proposed to conduct a research on issues related to the classical custom, history, politics, tradition, and religion of Ethiopia, he/she have to explore the works handed down from the previous generations to the current generation. So, he/she must investigate these literatures. In addition to this, as the language is the ancestor of other modern Ethio-Semitic languages like Tigrinya and Amharic (Marvin_Lionel_Bender, 1976), professionals of these languages should also know the linguistic nature of Ge'ez language to earnestly understand and investigate the nature of these modern ones. To use these resources, one must know the language itself or else these literatures have to be translated into either of the currently spoken languages manually, which may take a long time. To solve this problem studying the nature of the language computationally and finally releasing the resources out with the help of Information Technology (IT) to be used by everyone of this era is a critical assignment that deserves research. As the result, it is worth conducting research as to develop a part of speech tagger for Ge'ez to contribute to the complete usage of the language by the generation (Desta_Berihu, November, 2010).

## 3. STATEMENT OF THE PROBLEM

There are POS taggers that have been developed for international languages like English (Eric_Brill, 1992), Arabic (Hadni_Meryeme, et al., December 2013), Hebrew (Roy_Bar-Haim, et al., 1998), etc. and Ethiopian languages like Amharic (Binyam_Gebrekidan, 2009), Afaan-Oromo (Getachew_Mamo & Million_Meshesha, 2011), Kafi-Noonoo (Zelalem_Mekuria & Yaregal_Assabie, 2013), Tigrigna (Yemane_Keleta, et al., July 2016) etc. In general, rule based approach, probabilistic or stochastic approach and hybrid approaches which is the combination of stochastic tagging techniques such as hidden Markov models (HMMs) and rule based tagging techniques are used to develop POS tagger for various languages. However, the way they are applied depends on the characteristics of languages. As a result, these POS taggers cannot be applied directly for Ge'ez language. On the other hand, to our best knowledge there is no research conducted on POS tagger development for Ge'ez language which becoming a barrier for research and development works on higher level NLP applications. Hence, the absence of POS tagger system limits researches concerning the NLP of

Ge'ez language such as parsing (syntactic and semantic), machine translation, sentence grammar checker, spell checker, speech synthesizer etc. as it is used as a pre-processing component for the aforementioned NLP applications. Hence, conducting research on developing an automatic POS tagger for Ge'ez language worth paramount significance. The general objective of this research work is to develop POS tagger for Ge'ez language to solve these problems.

## 4. RELATED WORK

Literature review has a vital role for identifying the component of part of speech tagger, comparison of the approaches, detail understanding of problems, finding gaps, identifying methodologies, etc. In addition, in order to understand the problem books, articles, journals and other publications will be reviewed. So far, there is no readymade tag set for Ge'ez language that can be used in this paper. Hence, there are to construct a new corpus and discussions will make with the language experts in order to set tag sets for this language. A lot of POS tagging works were conducted in Semitic languages which are categorized under the same language branch as the Ge'ez language and other Ethiopian and international languages which are not Semitic language families.

**Semitic languages**

Yemane Keleta and Yamamoto Kazuhide (Yemane_Keleta, et al., July 2016) presents part of speech tagging research for Tigrinya from the newly constructed Nagaoka Tigrinya Corpus. The raw text was extracted from a newspaper published in Eritrea in the Tigrinya language. The POS tagged corpus contains 72,080 tokens and 73 tag set. Subsequently, a supervised learning approach based on conditional random fields (CRFs) and support vector machines (SVMs) was applied, trained over contextual features of words and POS tags, morphological patterns, and affixes. For a reduced tag set of 20 tags, an overall accuracy of 90.89% was obtained on a stratified 10-fold cross validation. Enriching contextual features with morphological and affix features improved performance up to 41.01 percentage point, which is significant.

Teklay Gebregzabiher (Teklay_Gebregzabiher, November, 2010) introduced a hybrid approach POS tagger for Tigrigna language. In this work the author used a combination of HMM, which is widely used under stochastic approach and adapting Brill transformation-error driven learning approach to drive machine learned rules for designing the rube based tagger component. The author has collected a total of 26,000 words from Tigrigna news broadcasting agencies and annotate manually with their corresponding word class. In addition to this, the author has identified 36 tag sets for the entire tagging process. Among the total word, 75% (20,000) words used for training purpose while the remaining 25% (6000) words used for testing purpose. Generally, this study finds tag of a word in two main steps. The first step is performed by the HMM tagger. The HMM tagger first annotates the given raw text and provides a level of confidence (threshold value) for each tag sequences. In the second step, the confidence level of each tag sequence compared with the minimum confidence level that is set by the author using the output analyzer module. In order to test the accuracy of the proposed method, the author conducted different experiment for the three types of taggers namely HMM tagger, rule based tagger and hybrid tagger. As a result, the author has got an accuracy of 89.13% for HMM, 91.8% for rule based and 95.88% for hybrid tagger.

Binyam Gebrekidan (Binyam_Gebrekidan, 2009) developed the POS Tagger for Amharic language. The author designed a POS tagger state-of-the-art machine learning algorithms for Amharic language. The author uses annotated data available for their experiments which is WIC corpus (≈207k) tokens. In order to increase the performance of the tagger the author uses the following three methods: First, the POS tagged corpus (WIC) has been cleaned up to minimize the preexisting tagging errors and inconsistencies. Second, the vowel patterns and the roots, which are characteristics of Semitic languages, have been used to serve as important elements of the feature set. Third, state-of-the-art of machine learning algorithms have been used and parameter tuning has been done whenever necessary and as much as possible. Finally, the accuracies have crossed above the 90% limit.

Hadni Meryeme *et al.* (Hadni_Meryeme, et al., December 2013) proposes POS Tagging technique for Arabic language using hybrid approach. The developed tagger employed an approach that combines rule-based method with HMMs based on the Arabic sentence structure. The proposed technique uses different contextual information of the words with a variety of the features which are helpful to predict the various POS classes. To evaluate its accuracy, the proposed method has been trained and tested with two corpora: The Holy Quran corpus and Kalimat corpus for discretized Classical Arabic language. Parts of it were used to train and to test the tagger. The experiment results demonstrate the efficiency of the method for Arabic POS Tagging. In fact, the obtained accuracies rates are 97.6%, 96.8% and 94.4% for respectively their Hybrid Tagger, HMM Tagger and for the Rule-Based Tagger with Holy Quran corpus. And for Kalimat corpus they obtained 94.60%, 97.40% and 98% respectively for rule-based tagger, HMM tagger and their hybrid tagger. In fact, the

accuracy was slightly increased with the increasing of the number of words in the training corpus. However, their tagger cannot handle for unknown words or tagging accuracy of unknown words ware very low. Additionally, their tagger cannot handle in extraction of multi-word terms.

**Non-Semitic languages**

Eric Brill (Eric_Brill, 1992) developed a simple rule-based tagger for English language with very few rules performs on par with stochastic taggers. The author ran two experiments where all words were known by the system. First, the Brown Corpus was divided into a training corpus of about one million words, a patch corpus of about 65,000 words and a test corpus of about 65,000 words. When tested on the test corpus, with lexical information derived solely from the training corpus, the error rate was 5%. Next, the same patches were used, but lexical information was gathered from the entire Brown Corpus. This reduced the error rate to 4.1%. Finally, the same experiment was run with lexical information gathered solely from the test corpus. This resulted in a 3.5% error rate. Note that the patches used in the two experiments with no unknown words were not the optimal patches for these tests, since they were derived from a corpus that contained unknown words.

Zelalem Mekuria and Yaregal Assabie (Zelalem_Mekuria & Yaregal_Assabie, 2013) developed the POS Tagger for Kafi-Noonoo using a hybrid approach. For training and testing purposes, 354 untagged Kafi-noonoo sentences are collected from two genres and annotated using an incremental corpus preparation approach. And 34 POS tags are identified for tagging purpose. After assigning word class information on each word within the sentences, both HMM and rule-based taggers are trained on 90% of the tagged sentences to generate probabilities i.e. lexical and transitional probability for the statistical component of the hybrid tagger and set of transformation rules for the rule-based component of the hybrid tagger. Based on these probabilities and transformation rules, the hybrid tagger assigns the most suitable word class information for the given untagged Kafi-noonoo texts. The performance of the prototypes i.e. HMM, rule-based and hybrid taggers ware tested using different experiments. As a result, HMM and rule-based tagger with unigram initial state tagger shows 77.19% and 61.88% accuracy respectively whereas, the hybrid tagger improves the accuracy to 80.47%. Even though there is no one way of choosing the size of training/testing set, this Paper applies heuristics such as 10% testing and 90% training corpus. But, doing so can bias the classification results and the results may not be generalizable.

Getachew Mamo and Million Meshesha (Getachew_Mamo & Million_Meshesha, 2011) presents part-of-speech tagger for Afaan-Oromo using HMM approach. For training and testing purpose, the authors collected 159 sentences (with a total of 1621 words) from different sources to make the corpus balanced and they used 17 tag set. In the tagging process, the tagger assigns word classes to a given Afaan-Oromo text with two main phases. In the first phase, the tagger trains on the training data in order to compute and store both lexical and transitional probability of training data. In the second phase, the tagger accepts untagged Afaan-Oromo text and tokenized into words. Then, the tagger assigns the correct POS tag for each token. This is achieved by using unigram and bigram model of the Viterbi algorithm by taking the stored information during the first phase. The authors have tested the performance of the tagger using tenfold cross validation mechanism. As a result, they have got 87.58% and 91.97% accuracy for unigram and bigram model respectively.

From the related work, researches can be done for both Semitic and non-Semitic language families that were conducted by different approaches such as hybrid approaches for Tigrigna (Teklay_Gebregzabiher, November, 2010), Arabic (Hadni_Meryeme, et al., December 2013), Kafi-Noonoo (Zelalem_Mekuria & Yaregal_Assabie, 2013), probabilistic approach for Afaan-Oromo (Getachew_Mamo & Million_Meshesha, 2011), Amharic (Binyam_Gebrekidan, 2009), Tigrigna (Yemane_Keleta, et al., July 2016) and rule based approach for English (Eric_Brill, 1992). For the best knowledge, there is no research attempt on Ge'ez POS tagger. As a result, the purpose of this proposed research is to fill in this research area gap.

## 5. SCOPE AND LIMITATIONS

The aim of this study is to develop POS tagger for Ge'ez words based on the corpus into their appropriate category. The corpus develop for this paper is domain specific corpus, a text corpus that will collect from a single domain, in this case the holy bible domain only. During the development of the corpus, the tag set use will have meant to give information of words about their word class category but not about the issues like gender, number, tense etc. Moreover, there are limited NLP researches done for Ge'ez language and hence there have been difficulties of using previous works as a reference. However, the text of the corpus will be written language words. Thus, this work is subject to the following scope and limitation:

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 1, pp: (12-23), Month: January - March 2020, Available at: www.researchpublish.com

- The tag set provide only word class information
- The corpus will be prepared from one genre that is holy bible.

## 6. POS TAGGING APPROACHES

There are different approaches to the problem of assigning each word of a text with a parts of speech tag, which is known as POS tagging (Fahim_Muhammad, et al., 2007) . The use of different tools and methods make the POS tagging different. At the top it can be supervised and non-supervised and used techniques to do it. The most common ones are rule-based, stochastic, artificial neural network and hybrid approaches. Figure1 demonstrates the classification of different POS tagging approaches (Fahim_Muhammad, et al., 2007) (Muhammad, et al., 2006) (Deepika_Kumawat & Vinesh_Jain, May 2015). This paper concerns on the supervised approach of POS tagging. The supervised POS tagging models require a pre-tagged corpus which is used for training to learn information about the tag-set, word-tag frequencies, rule sets etc. The performance of the models generally increases with the increase in size of the pre-tagged corpora (Muhammad, et al., 2006). The most commonly used supervised POS tagging are also TnT, TnT with Regex and hybrid of them.

In order to achieve the objectives of this research, relevant tools and methods are used. Python programming is selected to develop the system on the implementation phase as a programming language tool and hybrid approach as a method. Hybrid approach combines features of both the rule-based approach and statistical approach, the rule based approach and the Artificial Neural Network or other different two approaches. Like rule-based systems, they use rules to specify tags. Like stochastic systems, they use machine-learning to induce rules from a tagged training corpus automatically. The transformation based tagger or Brill tagger is an example of the hybrid approach. Most work on POS tagging have got better results than the corresponding uncombined approaches.

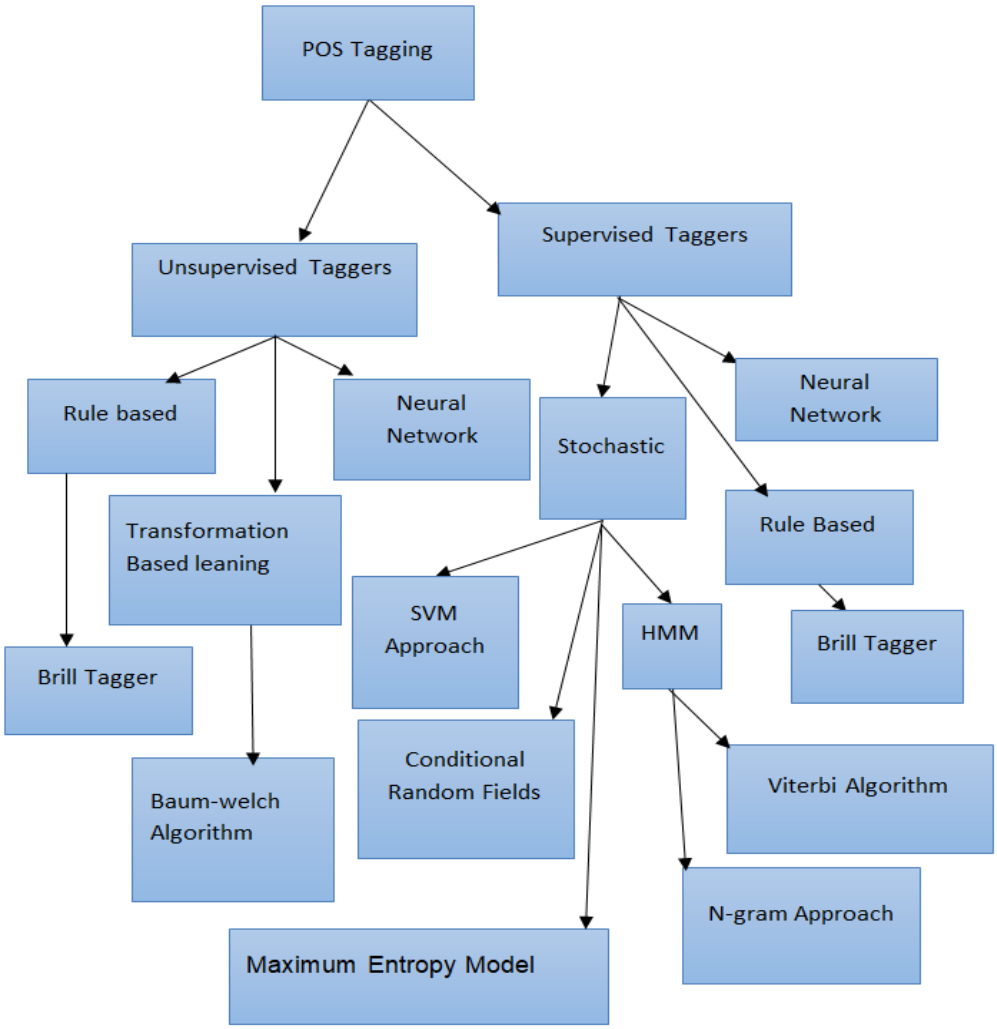


**Figure 1: Classification of POS tagging Approaches**

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 1, pp: (12-23), Month: January - March 2020, Available at: www.researchpublish.com

## 7. DESIGN OF THE GE'EZ POS TAGGER

POS tagging involves many difficult problems, such as insufficient amounts of training data, inherent POS ambiguities, and most seriously, many types of unknown words which are pervasive in any application and cause major tagging failures in many cases. In order to achieve the research objectives, relevant tools and approaches were used. The system is developed using python as a programming language demand for implementation extend from the possibility to combine the execution of other related components.

Several approaches have been proposed to annotate words automatically with their POS tags. Among these, the hybrid of TnT and rule-based approach is assumed to perform better than the TnT and rule-based taggers when they are taken alone. For this Paper, a hybrid approach, which is a combination of TnT, human annotated rule, and morphological pattern analyzer tagger is designed for Ge'ez language. The hybrid tagger of Ge'ez consists of three main components these are initial state (TnT tagger), output analyzer and rule-based tagger and morphological analyzer based tagger. TnT tagger associates the sequence of returned tags with the correct words in the input sequence.

The tester part, first untagged the golden standard sentences using manual tagging then gets a list of untagged and tagged sentences for testing. Next, it scores the efficiency of the tagger against the gold tag[1] standard. In other words, it strips the tags from the manually tagged standard text, retag it using the tagger, and then compute the accuracy score. The overall architecture of the system including the connection between the components and the algorithm are shown in Figure 2 and Figure 3 respectively.

Algorithm for Hybrid Tagger

1. Read the text to be tagged
2. Shuffle the input text
3. Segment and tokenize the sentence
4. Prepare unique words Dictionary
5. Prepare common prefixes Dictionary
6. Get trained TnT and Regex tagger
7. Tag the test corpus by using TnT and Regex
8. Compare with gold tag
9. While testing! = gold tag
   9.1. Apply the word with the given rule if the word contains wuetu
   9.2. Analyze morphological patterns of the word and tag accordingly
10. End of hybrid tagger

**Figure 2: Hybrid Tagger Algorithm**

The pre-processing component of hybrid tagger are have three main modules, sentence segmentation, word tokenizer and sentence shuffler module. The sentence splitter module accepts tagged texts using Ge'ez corpus reader and splits down at sentence level based on Ge'ez sentence end marker characters. The default sentence tokenizer is an instance of NLTK tokenize with '\n' to identify the gaps of sentences. It assumes that each sentence is on a line all by itself, and individual sentences do not have line breaks. The preliminary tagging component is the component that have very few manually tagged corpus.
This paper work customize this by passing in its own tokenizer to the function to tokenize Ge'ez language sentences.

[1] Gold tag is the set of manually tagged corpus

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 1, pp: (12-23), Month: January - March 2020, Available at: www.researchpublish.com

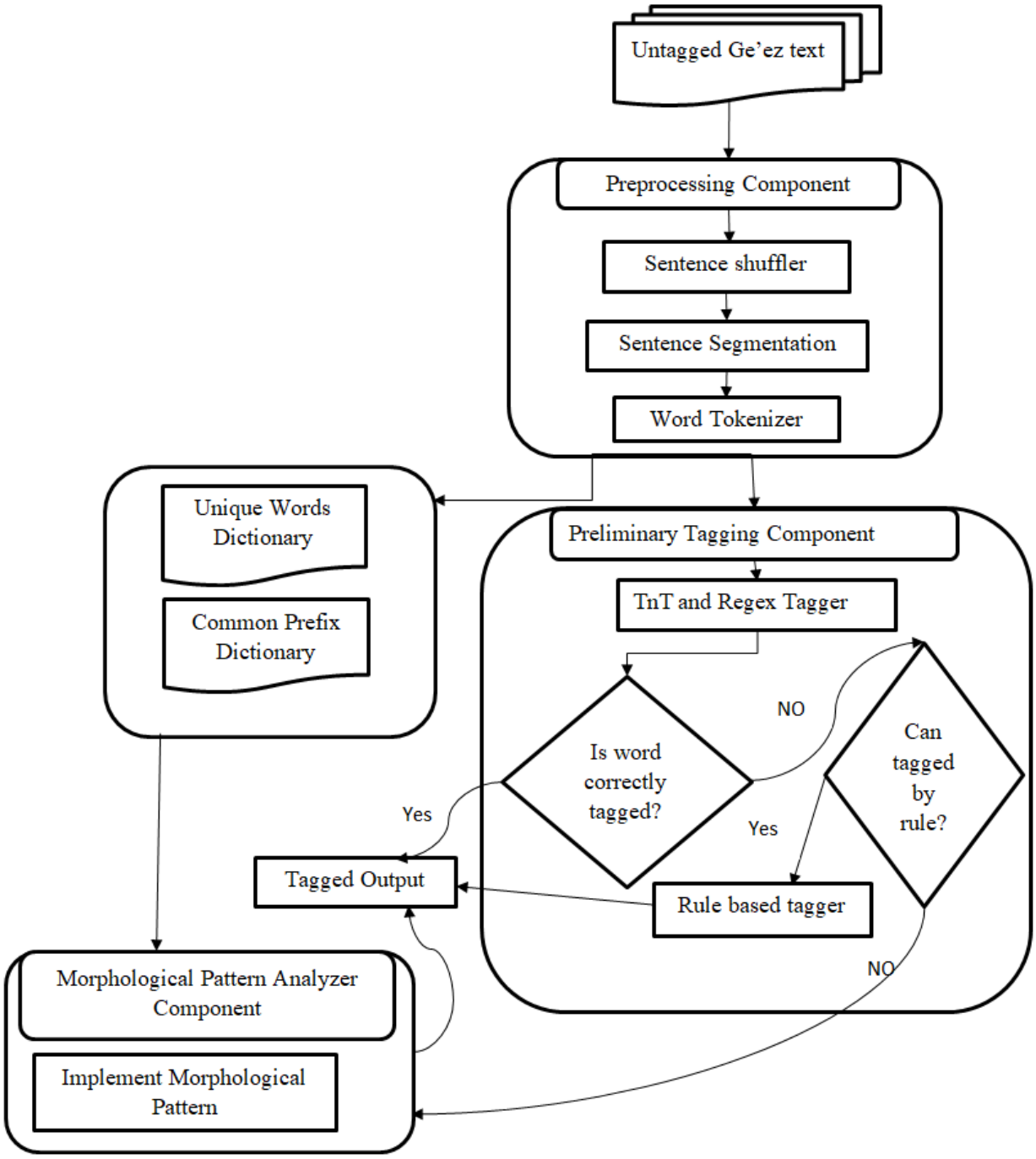


**Figure 3: Hybrid tagger architecture**

## 8. EXPERIMENT AND RESULTS

Several experiments with different training set on three POS tagger have been conducted for Ge'ez POS tagger. The entire corpus shuffled and divided into two main sets: training set and testing set. Using the 10 fold cross validation (10 fold CV) evaluation technique, the training set covers 90% of the entire corpus, the remaining 10% of the corpus is used for testing purpose. The Accuracy, recall, precision, and F-score are the evaluation measures of the performance of the system. The prototype of the system was developed and tested with samples tagged sentences. The performance of the system is measured against the manually prepared corpus.

| |
|---|
| *Accuracy=Total Number of Correctly tagged/ Total Number of Testing tag* |

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 1, pp: (12-23), Month: January - March 2020, Available at: www.researchpublish.com

### 8.1. Test Result of TnT Tagger

Use NLTK tool for implementing the experiment of Ge'ez TnT tagger by a little bit modification. Ten different experiments are conducted on the TnT tagger using different portions of the training set to see the excellence of the training set based on the observation that can be made on the learning curve. Started training the system using the 10% of the training set. After the tagger is trained, its performance is measured on the testing set. Having got a low performance of the tagger trained on the 10% of the training set, kept on adding the training data by 10% until they got a desired performance of the tagger. Table 1 shows the different experiments conducted using different portions of the training set with the corresponding performance of the tagger.

**Table 1: TnT tagger performance**

| Training set | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| Performance (%) | 50.65 | 58.51 | 62.78 | 66.08 | 69.08 | 70.44 | 72.57 | 74.08 | 75.187 | 77.87 |
| Difference | 50.65 | 7.86 | 4.27 | 3.3 | 3 | 1.36 | 2.13 | 1.51 | 1.107 | 2.683 |

The TnT tagger in Figure 8.2, shows 77.87 % accuracy when all of the training data is used (100%). However, this result indicated the worst performance of an annotating Ge'ez corpora. This can be explained by two reasons. The first one is as TnT is a statistical tagger for training purpose it needs large corpus size but use small corpus size which makes it very difficult for stochastic taggers to create probability distribution to hold transitions between different states. The second reason can be from grammar order in Ge'ez sentences in which free grammar which is no agreement among subject-object-verb order.

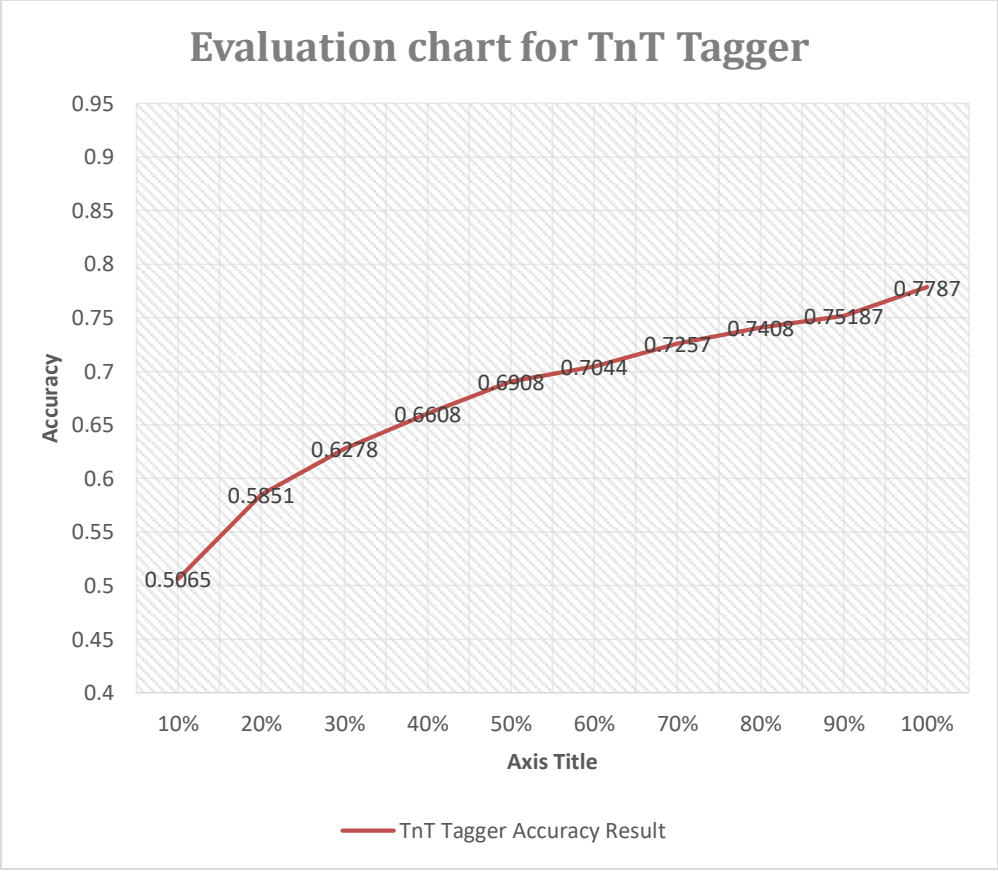


**Figure 4: TnT Tagger Performance curve**

Due to the aforementioned reasons, for Ge'ez language, TnT tagger score the worst accuracy result comparing with different language using this tagging approach, for example for English using Penn Treebank corpus which contains 50,000 sentences (1.2 million words) scores an accuracy of 96.7 % [23]. In the same manner (additionally) for Amharic language using 1065 news texts (210,000 words) score the overall performance 92% [45].

### 8.2. Test Result of TnT and Regex Tagger

To test the performance of the TnT and Regex Tagger like that of TnT tagger, ten different experiments are conducted using different portions of the training set. Table 6.2 shows the different experiments conducted using different portions of the training set with the corresponding performance of the TnT with back off of Regex tagger.

The most difficult task of TnT tagger is tagging of unknown words, words do not appear in training phase [23]. Hence, if the baseline TnT algorithm encounters a word in the testing set which did not appear in the training set, it will simply annotate it as "UNK" (unknown). Rather than failing to annotate in this way, the alternate versions of TnT identify a back off tagger. Thus, when the algorithm comes upon an unknown word, it will pass off the tagging task to the back off tagger. Such backoffs can be chained together but there is usually no additional improvement in having more than one or two backoffs. The most common class of lexeme in the corpus is nouns. TNT and Regex performs better than TnT based tagger. By replace "UNK" to "N" get a little bit accuracy change in the tagger. Figure 5, shows the curve shows 82.23 % which is 4.36 % difference comparing with TnT tagger in Figure 4.

**Table 2: TnT and Regex tagger performance**

| Training set | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| Performance (%) | 63.50 | 68.56 | 71.42 | 73.65 | 75.67 | 76.77 | 78.45 | 79.65 | 80.04 | 82.23 |
| Difference | 63.50 | 5.06 | 2.86 | 2.23 | 2.02 | 1.1 | 1.68 | 1.2 | 0.39 | 2.19 |

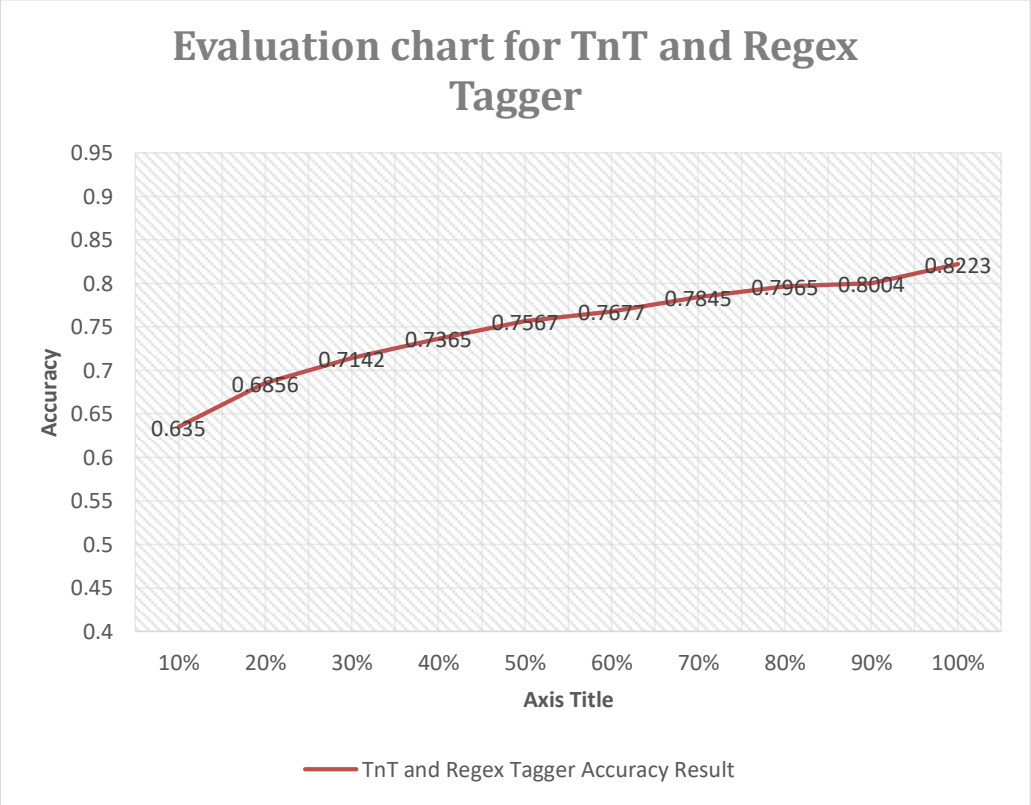


**Figure 5: TnT with Regex Tagger Performance curve**

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 8, Issue 1, pp: (12-23), Month: January - March 2020, Available at: www.researchpublish.com

### 8.3. Test Result of Hybrid Tagger

Hybrid tagger of Ge'ez language is combination of TnT with back off of Regex tagger and contains morphological pattern analysis. In order to tag a given text with the hybrid tagger, first the Regex assigns tags to tokens on the basis of matching patterns. For instance, guess that any word contains ten digits of numbers or match numbers with \d is a cardinal number, and is tagged as CR. It follows sequential order, and the first one that matches are applied. The final regular expression (r'.*', 'N'), is a catch-all that tags everything as a noun. The reaming task will be done by TnT tagger. Even though the combined tagger, TnT with back off of Regex tagger is better perform than TnT only, but still the result is acceptable. Consequently, it is important to associate morphological pattern analysis with the tagger which is making hybrid tagger. In addition to TnT with back off of Regex tagger, the hybrid tagger work by guessing unknown word using morphological pattern of the word. In unknown word guessing, the POS tag of an unknown word is predicted using affix of the unknown word, morphological patterns and substrings methods. Use probability method to guess the POS tag of unknown word. Finally, by combing all those techniques got an acceptable performance result.

**Table 3: Hybrid Tagger performance**

| Training set | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| Performance (%) | 87.24 | 89.46 | 90.46 | 91.30 | 91.97 | 92.24 | 92.85 | 93.40 | 93.92 | 94.32 |
| Difference | 87.24 | 2.22 | 1.00 | 0.84 | 0.67 | 0.28 | 0.61 | 0.55 | 0.52 | 0.40 |

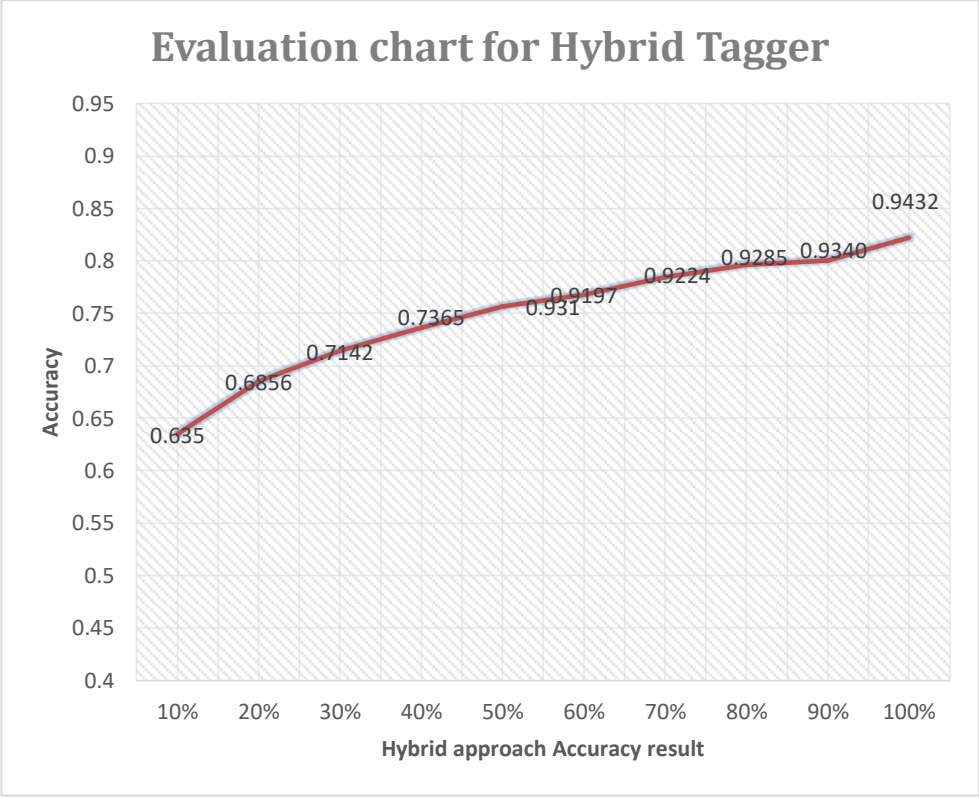


**Figure 6: Hybrid Tagger performance curve**

The Hybrid approach shows as 12.09% performance improvement as you can see from the Figure 6, shows the curve shows 94.32 % and this is the best than the individual taggers.

## 9. APPLICATION OF RESULTS

There are many advantages of developing POS tagger for a specific language. In the first place, it is the basis for developing other higher level applications of NLP such as parsing, information extraction, information retrieval, question answering, text to speech, etc. These applications can be used in different areas of the Ge'ez language. Accordingly, the beneficiaries of this study are:

- Researchers who want to conduct on higher level application of NLP for this language such as spell checker, grammar checker, speech recognition, etc.
- People who want to learn Ge'ez as a second language; it may help them to discover the word categories and grammar construction.
- It can be used as an input for full parser
- It can be used in text-to-speech system to correct the way of pronunciation
- It can be used for surface linguistic analysis

## 10. CONCLUSION AND FUTURE WORK

This paper provides Ge'ez POS tagging that is the process of assigning POS like noun, verb, preposition, pronoun, adverb, adjective or other lexical class markers to each word in a sentence or literature. POS tagging is the first step to understanding a natural language. Most other tasks and applications heavily depend on it. POS tagging, the research area in the field of NLP for different languages, is considered as one of the basic necessary tools. Several techniques have been suggested to tag words automatically with their POS tags. Among these, the hybrid of TnT with human annotated rule, regex and morphological pattern analysis of Ge'ez language is assumed to perform better than the TnT taggers taken alone.

Corpus is an important component in NLP in general and POS in particular. For this Paper, a corpus with a total of 1305 sentences is collected from one genre. For this Paper, 26 POS tags are identified as a tag set for annotating a raw text. The tag set indicates only word class rather than gender, number, tenses etc. The training set consists 90% of the total corpus (around 1175 sentences) and the testing set consists 10% of the corpus (around 130 sentences) using the 10 fold cross validation method via NLTK and Python3.6.2 as tools used in the implementation and experiment of the Ge'ez POS tagger. Hence, different experiments are conducted for the three types of taggers namely the TnT tagger, TnT with Regex tagger and Hybrid tagger. The results are 77.87%, 82.23% and 94.32% performances are obtained for TnT tagger, TnT with Regex tagger and Hybrid taggers respectively. Therefore, it is possible to conclude that the hybrid tagger performs better than the individual approaches of TnT tagger and TnT with Regex tagger for Ge'ez language and other Semitic languages.

Finally, this research work suggests the following key points as a future work:

- Preparation of a balanced corpus that contains texts which represent different genres like theological and hymn books such as Synaxarium (the book of the saints of the Ethiopian Orthodox Church), deeds of the martyrs etc. and other books beyond religious scriptures such as fictions, textbook etc.
- Comparative study of three different approaches (CRF, SVM classifiers based and ANN based taggers for Ge'ez Language with more training and testing data)
- Extending this work by training in large corpus and using large tag sets that can identify gender, number, tense etc. with different feature set
- Comparison of two hybrid approaches: the hybrid of ANN and TnT tagger and the hybrid of TnT and CRF for Ge'ez language
- Morphological pattern analysis component of hybrid approach that proposed for Ge'ez POS tagger is based on unknown word guessing mechanism. Therefore, in order to further improve the tagging results, this approach can be extended to use the full feature of Ge'ez morphological analyzer.

## REFERENCES

[1] Allen_James, 1995. *Natural language understanding.* CA, USA: Benjamin-Cummings Publishing Co., Inc..

[2] Anon.,n.d. *Chomsky-Definition.* [Online] Available at: https://www.scribd.com/doc/22325162/Chomsky-Definition [Accessed 02 11 2016].

[3] Binyam_Gebrekidan, 2009. *Part of Speech Tagging for Amharic.* UNITED KINGDOM: s.n.

[4] Deepika_Kumawat & Vinesh_Jain, May 2015. POS Tagging Approaches: A Comparison. *International Journal of Computer Applications (0975 – 8887),* Volume Volume 118 – No. .

[5] Desta_Berihu, November, 2010. *DESIGN AND IMPLEMENTATION OF AUTOMATIC MORPHOLOGICAL ANALYZER FOR GE'EZ VERBS.* s.l.:Unpublished.

[6] Eric_Brill, 1992. *A SIMPLE RULE-BASED PART OF SPEECH TAGGER.* Trento, Italy, s.n.

[7] Fahim_Muhammad, Naushad_UzZaman & Mumit_Khan, 2007. *Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla.* Bangladesh: Springer Netherlands.

[8] Getachew_Mamo & Million_Meshesha, 2011. Parts of Speech Tagging for Afaan Oromo. *International Journal of Advanced Computer Science and Applications,* Issue 2011.010301.

[9] Hadni_Meryeme, Ouatik_Said_Alaoui, Lachkar_Abdelmonaime & Meknassi_Mohammed, December 2013. Hybrid Part-Of-Speech Tagger for Non-Vocalized Arabic Text. *International Journal on Natural Language Computing (IJNLC),* Volume Vol. 2, p. No.6.

[10] Jurafky & Martin, 2009. Speech and Language Processing. In: *An introduction to natural Language Processing, Computational Linguistics, and speech recognition.*. 2nd ed. New Jersey: Prentice Hall.

[11] Leslau & Wolf, 1987. *Comparative Dictionary of Ge'ez (Classical Ethiopic).* Wiesbaden: Harrassowitz..

[12] Liddy_Elizabeth, 2001. Natural Language Processing. In: *Encyclopedia of Library and Information Science.* 2nd ed. New York: Marcel Decker.

[13] MahibreKidusanResearchCenter, 2010. Ethiopian church studies,. *Journal of Ethiopian church studies,the Ethiopian Orthodox Tewahido church Sunday schools department.*

[14] Marvin_Lionel_Bender, 1976. Language in Ethiopia. In: London: Oxford University Press, pp. pages 23-27 ; 99-106.

[15] Muhammad, F., Khan, MumitUzZaman & Naushad, 2006. *Comparison of different POS tagging techniques for some South Asian languages,* 2006: BRAC University.

[16] Roy_Bar-Haim, Khalil_Sima'an & Yoad_Winter, 1998. Part-Of-Speech Tagging of Modern Hebrew Text. *Natural Language Engineering,* 14(2), pp. 223-251.

[17] Sisay_Fissaha, 2005. Part of Speech tagging for Amharic using Conditional Random Fields. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages,* June , Issue Association for Computational Linguistics, p. pages 47–54.

[18] Teklay_Gebregzabiher, November, 2010. *PART OF SPEECH TAGGER FOR TIGRIGNA LANGUAGE,* Addis Ababa: Addis Ababa University.

[19] Yemane_Keleta, Yamamoto_Kazuhide & Marasinghe_Ashuboda, July 2016. Tigrinya Part-of-Speech Tagging with Morphological Patterns and the New Nagaoka Tigrinya Corpus.. *International Journal of Computer Applications,* Volume 146(14), pp. 33-41.

[20] Zelalem_Mekuria & Yaregal_Assabie, 2013. A Hybrid Approach to the Development of Part-of-Speech Tagger for Kafi-noonoo Text. November, Issue Unpuplished.