Date of publication DEC-31, 2021, date of current version AUG-25, 2021.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.20.4.2448

A Hidden Markov Model-based Part of Speech Tagger for Shekki'noono Language

ALEBACHEW CHICHE[1], HIWOT KADI[1], TIBEBU BEKELE[2]

[1]Debre Berhan University, College of Computing, Debre Berhan, Ethiopia (e-mail: alebachew.chz@gmail.com, hiwotkd73@gmail.com)
[2]Mizan Tepi University, School of Computing and Informatics, Department of Information Systems, Tepi, Ethiopia (e-mail: tibebu99@gmail.com)

Corresponding author: Alebachew Chiche (e-mail: alebachew.chz@gmail.com).

ABSTRACT Natural language processing plays a great role in providing an interface for human-computer communication. It enables people to talk with the computer in their formal language rather than machine language. This study aims at presenting a Part of speech tagger that can assign word class to words in a given paragraph sentence. Some of the researchers developed parts of speech taggers for different languages such as English Amharic, Afan Oromo, Tigrigna, etc. On the other hand, many other languages do not have POS taggers like Shekki'noono language. POS tagger is incorporated in most natural language processing tools like machine translation, information extraction as a basic component. So, it is compulsory to develop a part of speech tagger for languages then it is possible to work with an advanced natural language application. Because those applications enhance machine to machine, machine to human, and human to human communications. Although, one language POS tagger cannot be directly applied for other languages POS tagger. With the purpose for developing the Shekki'noono POS tagger, we have used the stochastic Hidden Markov Model. For the study, we have used 1500 sentences collected from different sources such as newspapers (which includes social, economic, and political aspects), modules, textbooks, Radio Programs, and bulletins. The collected sentences are labeled by language experts with their appropriate parts of speech for each word. With the experiments carried out, the part of speech tagger is trained on the training sets using Hidden Markov model. As experiments showed, HMM based POS tagging has achieved 92.77 % accuracy for Shekki'noono. And the POS tagger model is compared with the previous experiments in related works using HMM. As a future work, the proposed approaches can be utilized to perform an evaluation on a larger corpus.

KEYWORDS Parts of speech tagger; HMM; NLP; Shekki'noono language; Bigram.

I. INTRODUCTION

ETHIOPIA is a multi-lingual country which includes more than 85 nations and nationalities with morphologically, semantically, syntactically, and lexically different languages [1]. And Shekk'noono is one of the languages which serves as an official language in education. Natural Language Processing (NLP) is a branch of computational linguistics that is concerned with automated, computer processing of natural languages such as speech acts or texts. It deals with the processing and understanding of natural language using computers. According to the authors of [2-4], it performs useful tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech. It is also a medium for communication that is incorporated by every human being. It has many applications including machine translation, speech recognition, question answering, information retrieval system, and parts of speech tagging.

The process of assigning Parts of Speech for every word in a given sentence according to the context is called Parts of Speech tagging [1, 3, 4]. It is one of the useful tasks in

Natural Language Processing (NLP). It plays an important role in Speech and NLP such as Speech Recognition, Speech Synthesis, Information Retrieval, word sense disambiguation, and machine translation [5, 6]. Parts of speech ambiguity is a common feature of a majority of the world's languages. Many languages, especially on the African continent, are under-resourced in that they have very few computational linguistic tools or corpora (such as lexica, taggers, parsers, or tree-banks) available. It is a process of assigning a tag to every word in a sentence that serves as a preliminary task for carrying out tasks like chunking, dependency parsing, and named-entity recognition on any language. All of these NLP systems must use part of speech tagger as their preprocessor components for their best performance [6, 7].

According to [4, 9, 10], many languages, especially languages of developing countries, have no sufficient resources and tools required for the implementation of human language technologies. These languages are commonly referred to as under-resourced or pi languages. Shekki'noono language is among the under-resourced languages. Since the structural complexities of natural languages differ from each other, natural language processing applications, there is a need to develop natural language processing applications for each language.

So our work focuses on carrying out POS tagging for Shekkinoono. Much of the research in POS tagging was devoted to resource rich languages like English and French. African languages like Shekkinoono have received far too little attention. It is a very crucial task for any language processing activities [8]. Parts of speech tagging is one of the natural language core application areas and not yet that much developed for under-resourced languages like Shekki'noono language.

The Viterbi algorithm has been used for implementing the tagger and it is a dynamic programming algorithm that optimizes the tagging of a sequence, making the tagging much more efficient in both response time and memory consumptions of the corpus during training and testing. HMM is one of the statistical approaches in natural language processing activities and can tag more than 100 sentences in one second. It takes the product of two pieces of information together that is, it finds the tag sequence that maximizes the likelihood of the product of word probability (P (word /tag)) and tag sequence probability (P (tag / previous n tags)) [5].

The basic problem while doing this research work was the lack of data corpus and language experts needed for building and testing the model. The tag sets used are meant to give information of words about their word class categories only, but not about the issues like gender, number, and tense aspects, etc. The experimental training is mainly concentrated on bigram to know the performance of the tagger.

Given this circumstance, there is a need to develop a POS tagger for Shekkinoono. In this paper, we present an effective POS tagger using HMM machine learning approach for under-resourced Shekkinoono language. Finally, the POS tagger model is compared with the previous experiments in related works using HMM.

The findings of this study will have significance to initiate other researchers to participate in different computational researches of Shekkacho language since no research has been done in the area of computational linguistic in this language.

The paper is organized as follows: Section 2 provides related works on POS tagging based on literature for various languages. Section 3 describes the data preparation and language characterstics. Section 4 discusses methods and techniques used in this study. Section 5 presents the lexical analysis of the data. Section 6 describes the experimental results. Section 7 provides the comparision of the proposed work with othres. Section 8 concludes this paper.

II. RELATED WORKS

Although, it was proposed to develop POS tagging model for resource rich languages like English and French in many research works [1-20], much less attention was given to under-resource languages like Shekkinoono. And several methodologies were explored to develop part of speech tagging for Shekki'noono language [5, 11-15]. All of the sources which have been explored are to support as input for the Shekki'noono language tagger.

Yemane, et.al. [16] designed a Tigrinya part-of-speech tagging with morphological patterns and the new Nagaoka Tigrinya corpus. The researchers tried to collect from the newly constructed Nagaoka Tigrinya Corpus. The corpus was collected from a newspaper published in Eritrea in the Tigrinya language sentences to develop a POS tagger. From this 85 % of the annotated corpus is used for training and the remaining 15 % of the corpus is used for testing the tagger. For this study, 13 tagsets are deployed for developing the Tigrinya POS tagger. Moreover, the study has followed three steps to reach a final decision, namely statistical analyzer, tagset finder, and part of speech tagging determiner. Finally, the POS tagger was tested using different testing parameters. Based on these the tagger scored 82.26 % accuracy and 66.73 % accuracy using statistical data and without statistical data respectively. The other author called Lanka and Science [17], used the same method, the Hidden Markov Model approach, to develop a part of speech tagger for the Sinhala language. Similarly, the researchers used 2754 sentences with 26 tagsets to develop the Sinhala language part of speech tagger. In this study, two basic steps have been followed to train and test the part of speech tagger. The first step is training in which the training data is trained using the Viterbi algorithm under the N-gram model. The second step is the testing phase in which untagged row texts are passed through the trained algorithm trained on an annotated corpus. The performance of the POS tagger is tested using cross-validation evaluation mode. Based on cross-validation testing mode, the POS tagger registered an accuracy of 95 % for unknown texts. The authors in [12] also used Hidden Markov Model for developing part of speech tagging for the Manipuri language. The POS tagging process of the tagger

uses both unigram and bigram model for training the training corpus. In this study, the researchers have used 2000 tagged sentences for training with 97 tagsets. On this training set with 97 tagsets, the POS tagger has performed 92 % accuracy. On the other hand, the tagger performed 80 % accuracy with 12 tagsets. As stated by researchers, the performance of the tagger developed in this study is overweighed the previous POS tagger developed for the Manipuri language.

Siraj [11] developed the SomaliPart of Speech Tagger using machine learning Approach. The researcher developed a POS tagger using different machine learning approaches (i.e., HMM and CRF) and neural network model. Our Somali POS tagger outperforms the state-of-the-art POS tagger by 87.51% on a tenfold cross-validation. Moreover, the researcher also used relatively better Amharic tagsets and a large size corpus than the previous works for the Amharic Language. He used 30 tagsets and 210,000 words of text corpus for the experiment. The other researcher called Getachew [5] deployed the Hidden Markov Model to develop simple Parts of speech tagger. He defined 25 tagsets for the POS system. Lexical probability and contextual probability was used to find the most probable tag of a word. Yemane [18] proposed a machine learning approaches for amharic parts-of-speech tagging. Their aim was to improve the performance of POS tagging for the Amharic language. They used an extension of the existing annotated data, morphological knowledge, feature extraction, applying grid search by tuning parameter and the tagging algorithms were also examined to get a significant performance difference in comparison with the previous work. The newly constructed and extended tagsets have contributed to the high performance of the proposed approach. The researcher extended ELRC tagset and constructed a new tagset from Quran and Bible. The experimenatal results show that the average accuracy of 86.44, 92.27 ,95.87 for ELRC, ELRCQB and ELEC-Extended tagsets respectively. Lisette G., et al [8] described a morphological tagger for Spanish based on Cuban corpora using different tagging methods. The proposed tagger combines methods like Hidden Markov Models with some heuristics and dictionaries to come with better part-of-speech tagger. A morphological analyzer has been used to reduce possible grammatical tags and to obtain its morphological information also. The proposed tagger achieves 97.76 % accuracy for a given corpus.

Diesner [19] presented a part of speech tagger for the Arabic language. As stated by the authors, because of the absence of the readymade Arabic corpus, they tried to develop the Arabic POS tagger from the beginning. Similar to the POS tagger developed by Abate and Techibel [14] a hybrid approach incorporates rule-based and statistical methods to develop Arabic POS tagger. Gamback investigated detailed experiments using TnT SVM Tool and Mallet techniques on three different tagsets. From the experiment, the researchers received overall accuracies of 85.56%, 88.30%, and 87.87% for TnT, for SVM and MaxEnt, respectively, using the ELRC tagset.

From the above review, one can understand that though there are many works available for different languages still more works are expected to design better POS taggers for these languages. Herewith, developing the POS tagger is the ultimate solution to identify different types of tags and tagsets and assign word class to a given word in the given sentences accordingly to help the researchers to do further researches in the area of natural language processing.

Through this work, the research areas and directions in developing and deploying part of speech tagging are mentioned and the trends of POS tagging using the Hidden Markov Model (HMM) are studied. The proposed solution all over again diminishes the problems of investigating other language technology research which could be done on top of POS tagging.

III. SHEKK'NOONO CORPUS AND TAG SETS

A. CORPUS PREPARATION

The sources of sentences that are needed for parts of speech tagging for Shekki'noono language are diverse. After preprocessed those data sources are used for testing and training the performance of the Shekki'noon POS tagger. Such data sources in the Shekk'noon language can be find in newspapers. They include (Social, Economically, Political and Religious aspects), textbooks like (Primary and Secondary School), modules and proverbs. All of those sentences are collected in the form of soft and hard copies. For the work, we have used manual typing to change the hard copy to softcopy. The ways of tagging process are carried out manually with the help of an expert in linguistic field, who are currently working with the Textbook and curriculum development for the two languages. The Parts of Speech Tagging data will be divided into two subsets namely the training data set and test data set. The lexicon will be developed to calculate the probability of the listed word category from the developing corpus for the study. Subsequently, there is no tagset developed for Shekk'noon language for Natural Language Processing. Finally, we have used 1500 sentences (23,300 words) tagged corpus for training the tagger and evaluating its performance for Shekk'noon.

For this work, part of speech has been identified for Shekki'noon language. As indicated in [20], Shekki'noon language has all the lexical categories of words known to exist. These are nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, and interjections. These categories are characterized based on the lexical meanings of the words. So, the word class identification of a given word in a sentence should be based on the positional role of words for part-of-speech tagging. For instance,

Gabbiti dishe kero haggiye. 'Gabbito built a thatched house'.

Gabbiti dishe deeboo hammiye. 'Gabbito has gone to bring thatched'.

The word dishe'thatched' in the first sentence has taken the position of an adjective to describe the types of house,

whereas in the second sentence it has placed the position of a noun. So, even if the word is considered as a noun based on its lexical meaning, it can also be categorized in other categories based on its contextual position in the sentence.

As described in [20], in Shekki'noon the order of words in a sentence is subject-object-verb, commonly known as, subject-verb agreement. Let us illustrate it in the following sentence of Shekki'noono.

Gabbiti mit'o kut't'iye. 'Gabbito cut a tree.

In this sentence, Gabbito is a subject, mit'o 'tree' is an object, and kut't'iye 'cut' is a verb.

As indicated above, if words are not arranged in their appropriate places in a sentence, their messages will also be vague or will have no meaning at all. For instance, while the sentence *Amuuri bi nuuche iida waahane* 'Amuuri came with her friend' is correctly written following the order subject – object – verb of the language, the sentence iida waahane Amuuri bi nuuche 'with came Amuuri with a friend' written in the order verb-subject-object which do not follow the right order of words.

B. TAG SETS

As we have discussed before, for Shekk'noon language there is no work done in part of speech tagging. Because of this there is no identified and described tagset that can be directly used for such kinds of research works. However, identifying and describing the tagset for any language is tedious and time-consuming we have tried to identify and describe available tagset in Shekk'noon language.

Here we have tried to identify and describe 41 tags. In Table 1. sample tags are presented with their descriptions.

Table 1. Sample Tagset

No.	Tags	Word class
1	N	Noun
2	NPRE	noun including attached with preposition
3	VPREP	verb with preposition
4	NSUF	noun with suffixes
5	PRON	Pronoun
6	PRONSUF	pronoun with suffix
7	V	Verb
8	VSUF	Verb with suffix
9	ADJ	Adjective
10	ADJSUf	adjective with suffix
11	PRE	Preposition
12	Q	Questioning
13	PRONPRE	pronoun with preposition
14	VPRE	verb with preposition
15	ADV	Adverb
16	ADVPREP	adverb with preposition
17	CONJ	Conjunction
18	CN	Cardinal Number
19	ON	Ordinal Number
20	PUNC	punctuation

As indicated above, if words are not arranged in their appropriate places in a sentence, their messages will also be vague or will have no meaning at all. For instance, while the sentence *Amuuri bi nuuche iida waahane* 'Amuuri came with her friend' is correctly written following the order subject – object – verb of the language, the sentence iida waahane Amuuri bi nuuche 'with came Amuuri with a friend' written in the order verb-subject-object which do not follow the right order of words.

IV. Methods and Techniques

Much of the previous works in POS tagging has been focused to resource-rich languages like English, Arabic and French. But most of African languages like Shekki'noono have received too little attention so far. Among the most common under-resourced languages is Sheki'noon Language. This fact motivates us to implement a generic POS tagger for Shekki'noono. The significant contribution of this work is presented by: (1) standard corpus construction, (2) developing a POS Tagger, (3) comparing the proposed and related works. In addition, we release a new dataset of Shekki'noono language. In this work, we mainly used hidden Markov model for developing the POS tagger.

A. HIDDEN MARKOV MODEL

HMM approach was deployed for developing POS tagger since it does not need detail linguistic knowledge of the language like other methods such as rule based approach. Viterbi algorithm is used for HMM implementation. Commonly, the task of POS tagger developments starts with splitting the corpus into training and test sets. First, we train the HHM using the training set. And then we used the test set for evaluating the overall performance of the POS tagger. The architecture of the proposed approach is presented in Figure 1.

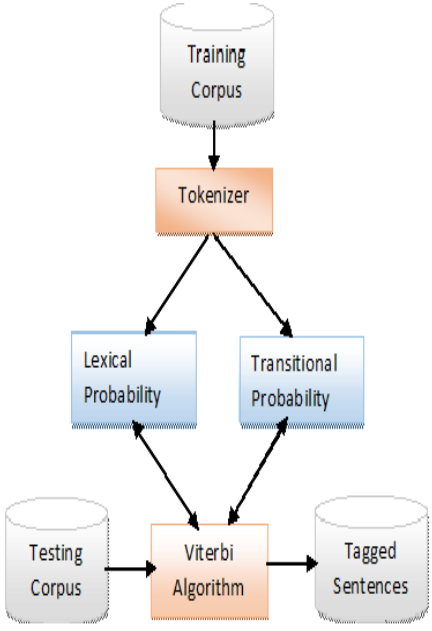

Figure 1 Architecture of the proposed POS tagger

V. Lexical Analysis

Lexicons are prepared for both the lexical and the transitional probabilities of the training corpus. From this, both lexical and transitional probability of each word is computed for each tag. Table 2 shows the lexicon disctribution of the corpus.

Table 2. Sample of lexicon distribution

	N	PRON	ADJ	ADV	V	CONJ	POS	Others	Total
Hetti	0	0	2	20	39	0	0	….	61
Boono	0	189	1	1	0	0	63	….	254
Bara	0	0	53	0	0	1	0	….	54
Tuna	0	0	2	2	9	0	0	….	13
Ariiye	5	0	26	0	1	0	0	….	32
Noono	22	5	0	0	0	0	0	….	27
Beeti	0	0	77	0	6	0	0	….	83
Kaamona	0	0	3	1	0	0	0	….	4
Toommoo	3	0	5	2	0	0	0	….	10
Wotta	0	0	0	3	1	22	0	….	26
...	…	...	…	…	…	…	…	…	…
Total	1377	710	3149	458	2091	191	293		8269

A. LEXICON PROBABILITY

As presented in Table 3 the lexicon probabilities of the tagger are computed by using a statistical method and it contains every word within the training corpus associated with its most frequent tag.

Table 3. Sample lexical probabilities

Words with their tags	Lexical Probabilities
P(hetti/V)	0.019
p(boono/PRON)	0.266197
p(tuna/V)	0.00430416
p(bara/ADJ)	0.0168307
p(noono/N)	0.0159767
p(wotta/CONJ)	0.11518324

For example, the lexical probability of the word "hetti" tagged with V within the lexicon distribution is computed as:

C (hetti, V) =24,

C (V) =611

So P (hetti/V) = C (hetti, V)/ C (V)

=24/611 =0.019

B. TRANSITIONAL ROBABILITIES

Shek'noon POS tagger deployed bigram to calculate the transitional probabilities of the lexicon corpus. So, the transitional probabilities, which are presented in Table 4, of the lexicon are computed by considering the information of the sequence of word-class category preceded by other categories. So, P (t\t-1) is used to compute transitional probabilities of lexicon corpus, where t is part of speech categories.

Table 4. Sample transition probability

Bigram categories	Probabilities
P(ADJ/N)	0.326
P(NSUF/CONJ)	0.042
P(CONJ/$)	0.11
P(CONJ/NUM)	0.033
P(CONJ/PRE)	0.068
P(VPRE/POS)	0.017
P(NPRE/PRE)	0.045
P(N/$)	0.082
P(NPRE/V)	0.019
P(ADv/ADJ)	0.03
P(N/PRE)	0.059
P(VSUF/ADJ)	0.014

For example, count (ADJ, N) =325 and count (N) =996.Therefore, the transitional probabilities for P (ADJ/N) =count (ADJ, N)/ count (N) =325/996 =0.326.

VI. EXPERIMENTS AND RESULTS

A. DATA

The data set used for training and evaluation is the manually tagged shekkonnno corpus. The Shekki'nono corpus contains 1500 senetences collected from over years from different domains. Every word within in the corpus is annotated with at least one tagset out of 41 possible handcrafted tags. It is implemented by splitting the data. For evaluation purpose of the tagger we have created a "standard" set of folds for the corpus. So, the "standard" folds are created by cutting the corpus into 10 pieces, each of the folds is about 1,500 words in sequence, while making sure that each piece contains full sentences (rather than cutting off the text in the middle of a sentence). Thus, the folds represent even splits over the corpus, to avoid tagging inconsistencies, but the sequences are still large enough to potentially make knowledge sources such as n-grams useful.

The resulting folds on average contain 23,066 tokens. Of those, 89.2% (20,574) are known, while 10.8% (2,491) are unknown, that is, tokens that are not in any of the other nine folds (if those were used for training, in a 10-fold evaluation mode). Then the experiments were conducted on the prepared folds. Here the performance of the tagger on the individual test sets has been calculated. To do so the count of words in the test set and the correctly tagged words in the standard folds were considered. Then the POS tagger was evaluated by matching the previously tagged output of the standard fold with the standard test set.

B. ALGORITHM FOR IMPLEMENTING HMM

For this work, the Viterbi algorithm of the hidden Markov model has been used to develop Shekk'noon POS tagger. The Viterbi algorithm needs the steps for searching and identifying one complete route The process continues till all

sequences of words in a sentence. The syntax is described as follows.

Let T = # of part-of-speech tags W = # of words in the sentence

for w = 2 to W

for t = 1 to T

Score (t, w) = P(Wordw/Tagt) * MAXj=1, T(Score(j, w-1) * P(Tagt/Tagj))

BackPtr(t, w) = index of j that gave the max.

This step is used for tagging processes for each word sequence depending on the information of variable *BackPtr*. It processes by iterating the *BackPtr variable* that holds the pointer of better category for each word squence in the sentence. The syntax of the process is described below.

Let T = number of part-of-speech tags W = number of words in the sentence

Seq(W) = t that maximizes Score(t, W)

for w = W-1 to 1

Seq(w) = BackPtr(Seq(w+1), w+1)

Zero(0) probabilities for transition and lexical are avoided by including the words that are newly faced during validation to the confusion matrix and tagging them as "UNK".

This indications to the lexical probabilities differ from zero for new words. Next to this, transition probabilities that comprise the UNK tag have not been seen previously and therefore equal zero(0), thus ending the propagation of the most likely paths through the pattern. To solve this problem, it is assigned a minimum probability (minprb) to the affected probability transitions: P (t | t=UNK) = minPrb and P(t=UNK | t) = minPrb.

Unknown (UNK) words are passed to a post-processing repetition that deploys a set of rules to re-annotate unknown (UNK) words with an appropriate part-of-speech. Following this technique, it is analyzed unknown (UNK) words that are received from multiple evaluation steps in order to get uniformities in their relationship with certain part-of-speech. On the basis of this analysis technique, we implemented four orthographic rules for tagging unknown (UNK) words. It is assumed that these rules are not created to be corpus-specific only, but also to be applicable for general use, for instance, words comprising a digit are to be tagged as numbers (D). Words which are capitalized are tagged as singular proper nouns (SPN). And also words that are ending in one out of 16 derivational or inflectional endings are to be tagged with the corresponding part-of-speech (for instance, words ending with "ing" are tagged as gerund (VG)). Finally, it is decided that every remaining unknown (UNK) word is annotated as noun (N). This result confirms the findings in [4, 11, 21], which were proposed for similiar language family like afaan Oromo and Somali language.

And the performance of the tagger was tested based on two ways. For experimenting HMM, the whole training corpus was segmented into ten equivalent sizes (each size is 10% of the total training data set). The performance of the POS tagger was evaluated by the first 10% of the data set and repeating the process by adding 10% of training data set to the previous data until the entire training corpus is used.

In the first experiment, the performance of the tagger is evaluated on the portion of the training data set. For the total data set 20% of the corpus is chopped for testing and 100% (including the test set) is used for training.

As a result, the performance of the POS tagger (Table 5) shows 78.05 % correctly tagged words out of the total 4513 words.

Table 5. Validating the Tagger with 20% Test Set

Correctly tagged Words	Incorrectly tagged words	Accuracy in Percent
3522	991	78.05%

In the second type of performance analysis, the tagger is repeatedly trained and tested following tenfold cross-validation as shown in the following examples. The following example shows how bigram tagger works.

Standard test set 1 is: **<s> /<s>** Ikka /ADJ nooneesse /NPOS **ditte /ADJ** wuro /N bari_bara /ADJ noononoshi /NSUF tune /V toyoo /NSUF boono /PRON bedigaa /VPRE noono /N shiijjaye /ADV hetiyaye /V. /PUNC </s> /</s>

Experimental result 1 is: <s> <s>ADJ Ikka NPOS nooneesse**V ditte**N wuroADJ bari_baraNSUF noononoshiV tuneNSUF toyooPRON boonoVPRE bedigaaN noonoADV shiijjayeV hetiyayePUNC. </s> </s>

From the above sentence of the standard test set, there is only one word tagged incorrectly. For example, the word "ditte" was tagged as "V" but the words must be tagged as "ADJ".

Standard test set 2 is: **<s> /<s>** Noononoshi /NSUF biini /PRON dittabeetonoshi /ADJ gogabeeti /VPRE beeye /N ikkinattonashona /NSUFPRE, /PUNC **hajje /ADJ** geedonoshi /NSUF kaamona /NPRE tuk'iyo /V **boonoshissi /POS** beetone /V. /PUNC **</s> /</s>**

Experimental result 2 is: <s> <s>NSUF NoononoshiPRON biiniADJ dittabeetonoshiVPRE gogabeetiN beeyeNSUFPRE ikkinattonashonaPUNC, **POShajje**NSUF geedonoshiNPRE kaamonaV tuk'iyo**PRONPOS boonoshissi**V beetonePUNC. </s> </s>

The experimental result of standard test 2 shows that the word "hajje" is tagged as "POS", the word "boonoshissi" is tagged as "PRONPOS" but the word "hajje" must be tagged as "ADJ" and word "boonoshissi" as "POS" respectively.

To determine the performance of the tagger, we need to count the number of words which can be tagged correctly and incorrectly by the tagger.

As a result, the performance of the bigram tagger is presented in Table 6 as follows:

Based on the experimental results depicted in the above Table 6, the bigram tagger gives better performance with an average accuracy of 91.28% with correctly tagged words. After all, the bigram tagger is selected as a best-performed tagger to implement the HMM POS tagger for Shekk'noon language. So, based on the tagging accuracy, we can

understand that the bigram tagger is a better performing tagger on the corpus prepared.

Table 6. The Bigram accuracy of Shek'noon POS tagger

Tested on	Count of words	Correctly tagged words	Incorrectly tagged words	Accuracy in percent (%)
Fold1	1524	1396	128	91.58
Fold 2	1387	1235	152	89.03
Fold 3	1423	1287	136	90.39
Fold 4	1728	1518	210	87.80
Fold 5	1472	1398	74	94.94
Fold 6	1766	1694	72	95.90
Fold 7	1392	1253	139	89.96
Fold 8	1456	1344	112	92.27
Fold 9	1654	1582	72	95.63
Fold 10	1348	1151	197	85.33
Accuracy				**91.28**

In general, different experiments were conducted for Shekki'noono HMM POS tagger. As a result, different performances are obtained from the experiments. To sum up, the Sheki'non POS tagger result shows that it assigns 13,858 (91.28 %) words correctly as tagged in the training set and 1,292 (8.82 %) words are tagged incorrectly different from the tag assigned in the training set. These performances are registered on the prepared testing set for Shekki'noono HMM POS taggers. A performance comparison for Shekki'noono HMM POS tagger indicates that it is better in terms of accuracy with the tenfold evaluation method. From the above performance analysis one can understand that if there are more data sets incorporated for training and testing, it may be possible to get higher performances than the archived performance results.

The proposed system has scored an average accuracy of 91.28%. As of our knowledge, the performance of this proposed system has presented a promising result even though we have used the smallest of the corpus that have already been tagged.

To our knowledge, the main reason for getting low performance on some of the tag sets are: (1) our model is not incorporated a rule based approach to detect abnormalities, (2) since the language is under resourced, no standard corpus available, and (3) manual tagging might be prone to error during labeling task.

VII. COMPARISION WITH PREVIOUS WORKS

To the best of our knowledge, no previous research works on POS tagging have been done for Shekkinono language. Therefore, we compared our proposed system with the prior POS-tagging model which is researched in language grouped into similar with Shekki'noono language family. Among these grouped languages, [4, 21] an Afaan Oromo POS tagger has been proposed using Hidden Markov Model (HMM) for Afaan Oromo language. The highest POS tagging accuracies have been achieved by HMM model. The HMM achieved an average accuracy of 90.77% on a tenfold cross-validation. On the other hand, a hybrid approach using Hidden Markov Model and rule based approach was proposed to develop a part of speech tagger for Kefinoono [15]. So, HMM tagger and rule based taggers are trained on 90 % the 354 sentences. The experimental result shows that Hybrid, HMM and rule based taggers achieved 80.7%, 77.19 and 61.88 % of accuracy respectively.

Also, the Somali POS tagger [11] was proposed using HMM, CRF, and neural networks for part-of-speech tagging for Somali language. The highest accuracies have been registered by all three HMM, CRF, and neural networks. All the three taggers achieved an average accuracy of 87.51 % on a tenfold cross-validation while under the same conditions.

Although we did not have any standardized and large size labeled corpus available for Shekki'noono, our proposed model outperforms better than proposed works for Somali, Kefinoono and Afaan Oromo with an average accuracy of 91.28%.

VIII. CONCLUSION AND FURTHER WORKS

Part of speech tagging is the process of assigning words within sentences with their corresponding part of speech or word categories. It is a hot research area in the field of natural language processing for different languages. Moreover, part of speech tagging can be conceived as the problem of assigning part of speech tags to a word in a sentence. This problem can be solved using different techniques among which the HMM-based approach is assumed to be one of the familiar approaches for implementing part of speech tagger. POS tagging is an initial stage of language technology, text analysis like information retrieval, machine translation, text to speech synthesis, information extraction, etc. The importance of the problem focuses on the fact that the POS is one of the first stages in the process performed by various natural language processing works. There are different approaches to address the problem of assigning a part of speech (POS) tag to each word of a sentence. Here, we have prepared for the Shekki'noono corpus and a tagset for the collected corpus with the help of linguistic experts. This work has implemented the Hidden Markov Model (HMM) to assign parts of speech. We have conducted experiments on the collected corpus using the bigram model. To conduct the experiments the corpus has been divided into two sets (training set and testing set) for testing and training purpose. The experiments tested using tenfold cross-validation evaluation mode attained an accuracy of 91.28% for Shekki'noono POS tagger. The second experiment was conducted using the percentage split evaluation technique by splitting the corpus into 80% training and 20% testing sets. In this experiment, the performance of the Shekki'noono POS taggers scores 78.05% Accuracy. So, the bigram tagger achieves better performances in tagging a given word using a tenfold cross-validation evaluation technique that checks the occurrences of words together with one word before.

Due to the limitations of the previous works and tagger explored in Shekki'noono language, the findings presented in this work can be used as a baseline for further research works in the area of parts of speech tagging. Further, tagging data with unknown words is also an essential need to handle in the tagger. When the system reaches an unknown word, the current tagger fails to propose a tag, since the system is not trained for that word and the tagging algorithm does not have enough intelligence to propose tags for untrained words and tags it as UNK.

In this work, a promising result is realized in developing Shekki'noono part of speech tagger for assigning a word classes to a given word. As a future research direction, the following issues are recommended. In this paper, we have been trying to present POS tagging for Sheki'noon using the HMM approach. In addition to exploring techniques other than HMM in POS tagging, it is also important to develop linguistic resources. There will not be major advances in Sheki'noon POS tagging the same stochastic methods unless the existing corpus is cleaned further or a new one is developed. This work can be extended by using more training and testing data and using a large tagset that can identify gender, number, tense, etc. by incorporating different features. And also it would be better to involve more language experts in manual labeling of the data. The present work is a beginning in Shekkinoono POS tagging. As a future work, the proposed approaches can be utilized to perform an evaluation on a larger corpus and carry out experiments using different methods for the existing corpus.

Data Availability

The dataset used in this work is prepared from scratch for research purposes. So, the processed data obtained to support the findings of this work is available from the authors upon request.

References

[1] B. A. Bilel and Y. Fethijarra, "Genetic approach tagging," *International Journal on Natural Language Computing (IJNLC)*, vol. 2, no. 3, pp. 1-12, 2013. https://doi.org/10.5121/ijnlc.2013.2301.

[2] K. Deepika and J. Vinesh, "POS tagging approaches: A comparison," *International Journal of Computer Applications*, vol. 118, no. 6, pp. 32-38, 2015. https://doi.org/10.5120/20752-3148.

[3] A. Tukur, K. Umar and S. A. S. Muhammad, "Parts-of-speech tagging of Hausa-based texts using hidden Markov model," *Dutse Journal of Pure and Applied Sciences (DUJOPAS)*, vol. 6, no. 2, pp. 303-313, 2020.

[4] M. Getachew, M. Million, "Parts of speech tagging for Afaan Oromo," *Int J Adv Comput Sci Appl*, 2015. https://doi.org/10.14569/SpecialIssue.2011.010301.

[5] Y. Getnet, *Unsupervised POS tagging for Amharic*, Master's Thesis, University of Gondar, Ethiopia, unpublished, 2015.

[6] J. Singh, N. Joshi and I. Mathur, "Part of speech tagging of Marathi text abstract using trigram method," *International Journal of Advanced Information Technology*, vol. 3, no. 2, pp. 35-41, 2013. https://doi.org/10.5121/ijait.2013.3203.

[7] D. Kumar, "Part of speech tagger for morphologically rich Indian languages: A survey," *International Journal of Computer Applications*, vol. 6, no. 5, pp. 1-9, 2010. https://doi.org/10.5120/1078-1409.

[8] G. Lisette, P. Aurora and R. Leonel, "A proposal of a morphological tagger for Spanish based on Cuban corpora," *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 210-214, 2018.

[9] Z. Fantahun, *Unsupervised Part of Speech Tagger for Amharic Language*, MSc. Thesis, Addis Ababa University, Addis Ababa, 2013.

[10] A. J. P. M. P. Jayaweera, and N. G. J. Dias, "Hidden Markov model based on art of speech tagger for Sinhala language," *International Journal on Natural Language Computing (IJNLC)*, vol. 3, no. 3, pp. 9-23, 2014. https://doi.org/10.5121/ijnlc.2014.3302.

[11] S. Mohammed, "Using machine learning to build POS tagger for under-resourced language: the case of Somali," *International Journal Information Technology*, vol. 12, pp. 717-729, 2020. https://doi.org/10.1007/s41870-020-00480-2.

[12] K. R. Singha, B. S. Purkayastha and K. D. Singha, "Part of speech tagging in Manipuri with hidden Markov model," *IJCSI International Journal of Computer Science*, vol. 9, no. 6, pp. 146-149, 2012.

[13] B. Gamback, "Tagging and verifying an Amharic news corpus," *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, 2012, pp. 79-84.

[14] S. T. Abate and M. Y. Tachbelie, "Designing and creation of pronunciation lexicons for speech processing in under-resourced and morphologically rich languages: The case of Amharic," Project Report of a Google supported research, 2014.

[15] Z. Mekuria, *Design and Development of Part-of-Speech Tagger for Kafi-Noonoo Language*, MSc. thesis, Addis Ababa University, School of Graduate Studies, College of Natural Sciences, Department of Computer Science, November, 2013. https://doi.org/10.1007/978-3-642-54906-9_17.

[16] K. Yemane, Y. Kazuhide, M. Ashuboda, "Tigrinya part-of-speech tagging with morphological patterns and the new Nagaoka Tigrinya corpus," *Int J Comput Appl*, vol. 146, no. 14, pp. 975–987, 2016. https://doi.org/10.5120/ijca2016910943.

[17] A. J. P. M. P. Jayaweera, and N.G.J. Dias, "Hidden Markov model based on art of speech tagger for Sinhala language," *International Journal on Natural Language Computing (IJNLC)*, vol. 3, no. 3, pp. 9-23, 2014. https://doi.org/10.5121/ijnlc.2014.3302.

[18] Gashaw and H. L. Shashirekha, "Machine learning approaches for amharic parts-of-speech tagging," *Proceedings of the ICON-2018*, Patiala, India, pp. 69–74, 2018.

[19] Diesn, "Part of speech tagging for English text data," School of Computer Science, Carnegie Mellon University, Unpublished, p. 1–8.

[20] Bechiro, D. Ambo, Shekki'noone Fiinniiyeessona shicheesse sheero, Maasha, Shekka: unpublished, version 3, 2017.

[21] T. Nedjo, D. Huang, X. Liu, "Automatic part-of-speech tagging for Oromo language using maximum entropy Markov model (MEMM)," *J Inf Comput Sci*, vol. 11, no. 10, pp. 3319–3334, 2014. https://doi.org/10.12733/jics20103906.

ALEBACHEW CH. ZEWDU was born in Merhabete, Shewa, Ethiopia. He has received his B.Sc. In Information Systems from Hawassa University in 2012 and M.Sc. In Computer Networking from Jimma University in 2016. He is currently working as a lecturer in the College of Computing, Debre Berhan University. His current research interests include Network Security, Data mining, and, Artificial intelligence, and data science. He has published more various research articles in these areas.

HIWOT KADI KUMSSA, received her B.Sc. Degree from Addis Ababa University, Ethiopia, in 2012, and M.Sc. from the University of Gondar, Ethiopia, in 2016. She is a lecturer at Debre Berhan University, Ethiopia. Her Research interests include Artificial Intelligence, Data Mining and Knowledge Discovery, Machine Learning.

TIBEBU BEKELE SHANA received his B.Sc. Degree from Jimma University, Ethiopia, in 2009, and M.Sc. from the Haromaya University, Ethiopia, in 2014. He is a lecturer at Mizan Tepi University, Ethiopia. Her Research interests include, Data Mining and Knowledge Discovery, Knowledge management, Information Science.