APPLICATION OF HYBRID APPROACH FOR WOLAITA LANGUAGE PART OF SPEECH TAGGING

Birhanesh Fikre
School of Informatics, Computer Science Wolaita sodo University, Ethiopia
fikrebire0916@gmail.com

ABSTRACT

The main purpose of this study is to develop part-of-speech tagger for Wolaita Language using hybrid approach. Part of speech tagger is one of the subtasks in NLP application which is important for other Natural Language Processing (NLP) applications, like parser, machine translator, speech recognizer and search engines. PoST is a process of tagging a corresponding part of speech tag for a word that tag defines how the word is used in a sentence. The PoST for Wolaita language is not enough yet to be used as one vital module in other natural language processing applications. In this study, the development of PoS tagger using hybrid approach that combines HMM and rule based approaches was conducted for Wolaita language. In general HMM model need large data to increase the performance and the rule based model learner rule based on the language features. The HMM tagger, tags the words based on the optimal path for a given sequence of words and transformation based learning (TBL) is a rule based approaches that learns rule directly from the training corpus without expert knowledge. The developed hybrid approach of Wolaita language PoS tagger uses HMM tagger as initial annotators and rule based tagger as a corrector based on fixed threshold values. For implementation and experiment purpose the researcher used python programming and NLTK. For training and testing the models, 1256 sentences or 15,268 words are collected from three different categories (Bible, Social media in Wolaita language (Wogetta FM 96.6) and Wolaita language department) and annotated data manually. For tagging purpose 26 PoS tag are identified. From entire corpus, 90% for training and the remaining of entire corpus for testing purpose. The performance of the taggers, are tested by using different experiments. After experiment the researcher found that the performance of HMM, rule based and hybrid taggers shows 88.14%, 93.19% and 94.82% respectively. Generally, hybrid approach showed the better performance to assigning part of speech tag for Wolaita language.

Key words: NLP, HMM, TBL, NLTK and Hybrid

1. INTRODUCTION

Natural language processing (NLP) is the ability of computer program to understand human language as it is spoken or written. It is the component of artificial intelligence (AI), which is automatic manipulation of natural language, like speech and text by a computer software. Natural language can be applied to any language that human beings use to communicate with each other [1]. NLP has many application areas. Some of these are text-to-speech and speech recognition, natural language dialogue interfaces to databases, information retrieval, information extraction, document classification, document image analysis, automatic summarization, spelling and grammar checking, machine translation, Part of Speech (PoS) tagging, plagiarism detection, stemming and others[2]. In this research our focus is on part-of-speech tagging for Wolaita language. Advanced natural language processing application requires part of speech tagging as preprocessing since identifying the part of speech or word class of a token is very important to determine its morphology, pronunciation and even semantics. Part of speech tagging is the preprocessing step of the advanced NLP application like text to speech, information retrieval, information extraction, parsing, machine translation and other NLP application. So, PoS tagging is vital for advanced NLP application, especially for languages that are under resourced. Likewise, for Wolaita language, to develop other advanced level NLP applications, PoS tagging is one of the fundamental NLP tasks that need to be accomplished. Wolaita language is one of the Northern Omotic languages that is spoken in the Wolaita Zone and some other parts of the Southern Nations, Nationalities, and People's Region of Ethiopia. It is also spoken in different cities of the country by the people from Wolaita and neighborhood Zones. The language has around 3.3 million native and dialective speakers[3]. However, computer applications that help to use the language in a more advanced way like text summarization, information retrieval and extraction, parsing and machine translation[2] are not available for this language. Works in natural language processing also showed that these high-level applications require low level language processing systems like part of speech tagging. The aim of this study, therefore, is to put an effort in preparing the groundwork in advance as part of speech tagging that enable to the next phase of natural language processing application developments. In Ethiopia, there are more than 80 different languages and among these languages, only four languages currently working with technological center especially in the telecom communication system. So, this study tries to contribute to the advancement to overcome the problems mentioned above. To use computers for understanding and manipulation of Wolaita language, there are few works conducted in this language. These tries include text-to-speech system for Wolaita language[4], speaker dependent speech recognition for Wolaita Language[5], development stemming algorithm for Wolaita language[6] and development of longest-match based stemmer for texts of Wolaita language[3]. There are also other related researches that were conducted on other local language. Especially on Amharic language, some researches were conducted on PoS tagging by[7] and[8], but in the Wolaita language there is a beginning work with few words in PoS tagging research developed[9]. Hence, it is the aim of this study to develop advanced automatic PoS tagger for Wolaita language to establish the base for future researchers who will have interest in the area of machine translation, information retrieval and text summarization.

2. Related Work

Development of Part of Speech Tagger Using Hybrid

Another early work for Afaan Oromo[10] is developed by Getachew Emiru using hybrid approach. In his work, he has developed part of speech tagger using hybrid approach that combines rule based and HMM approaches was conducted for Afaan Oromo. The transformation based learner, which is a rule based tagger, tag the words based on rules, or transformations induced directly from the training corpus without human intervention or expert knowledge. The HMM tagger, tags the words based on the most probable path for a given sequence of words. For implementation and experiment, he used NLTK 3.0.2 and python 3.4.3 and 1517 sentences were used for training and testing, from these sentences 85% for training and the remaining 15% for testing. The performance analysis of the three taggers, namely: HMM, rule based and hybrid tagger were tested with the same training and testing set they achieved accuracy of 91.9%, 96.4% and 98.3%, respectively. Based on their performance and learning curve analysis, the study has concluded that the hybrid tagger has been benefited from the advantages of the two separated approaches and achieved an improved performance.

Design and Development of Part of Speech Tagger for Kafi-noonoo Language

The researcher developed part of speech tagger for Kafi-noonoo language in the paper of [11]. The author developed tagger using hybrid approach i.e. HMM and rule based tagger at sentence level. He used 354 untagged Kafi-noonoo sentence are collected from two genres and annotated using an incremental corpus preparation approach for training and testing purpose. The researcher identified 34 part of speech tags for tagging purpose, for training 90% of tagged sentence are used. The performance of HMM, rule based and hybrid taggers are tested using different experiment. As result, the performance of HMM, rule based and hybrid tagger shows 77.19%, 61.88% and 80.47% accuracy respectively. The author concluded that the hybrid tagger outperform for Kafi-noonoo language.

Hybrid Part-of-Speech Tagger for Non-Vocalized Arabic Text

This research reported on the efficient and accurate part of speech tagging techniques for Arabic language using hybrid approach[12]. The HMM integrated with Arabic rule based method and to evaluate the accuracy of the proposed tagger, a series of experiment were conducted using holy Quran corpus and kalimat corpus for undiacritized classical Arabic language. The researchers used two corpuses to trained and tested PoS tagger for Arabic text. The authors was used the holy Quran corpus to evaluate PoS tagger , the evaluation rate are 97.6%, 96.8% and 94.4% for respectively hybrid tagger, HMM tagger and rule based tagger and the evaluation rate of kalimat corpus are 94.60%, 97.40% and 98% for respectively rule based tagger ,HMM tagger and hybrid tagger. The researchers concluded that accuracy of hybrid method represents a very good result compared with Tanni‟s rule based method and Albared‟s HMM method.

Part-of-Speech Tagger for Tigrigna Language

The research work in[13] was developed PoS tagger for Tigrigna by using hybrid approach, HMM tagger combined with rule based tagger for Tigrigna part of speech tagger. As result, for training and testing purposes the author identified 36 broad tag sets and 26,000 words from around 1000 sentence containing 8000 distinct words were tagged. The researcher was used the way of combining HMM with rule based, first tagged raw Tigrigna text by using HMM tagger; afterward the rule based tagger is used as a corrector of HMM tagger. In this work the researcher was use Viterbi algorithm and Brill transformation based error driven learning are adapted for the HMM and rule based taggers respectively. The corpus divide into training set and testing set, for training set 75% of corpus was used and for testing set 25% was used. The different experiments are conducted for the three types of tagger (rule based tagger, HMM tagger and hybrid tagger). Thus, 89.13%, 91.8% and 95.88% performances are attained for HMM tagger, rule based tagger and hybrid tagger respectively. As result, the researcher concluded that the hybrid tagger is better than HMM tagger and rule based tagger used individually.

Hybrid Approach for part of Speech Tagger for Hindi Language

In [14] the researchers proposed hybrid based part of speech tagger for Hindi language. This system is developed using the combination of HMM tagger and rule based tagger and for the experiment the authors used 80,000 words corpus with 7 different standard part of speech tags for Hindi. This proposed system works in two ways-firstly input words are found in database, if it is present then it is tagged. Secondly if it is not present then applied various rule or HMM model. As result, the authors concluded that the hybrid approach has good performance for PoS tagging.

Hybrid part of speech tagger for Malayalam

The research paper of[15] proposed an efficient and accurate PoS tagger for Malayalam language by using hybrid approach. Conditional Random Field (CRF) method integrated with rule based method and the researchers used SVM based method to compare the accuracy. Both tagged and untagged corpus used for training and testing the system and the proposed approach is Unicode based. As result, the authors presented the accuracy of 94% for PoS tagger of Malayalam.

Hybrid part of Speech Tagger for Sinhala Language

In the paper [16] the researchers developed part of speech tagger for Sinhala language using hybrid approaches. The authors combined stochastic with HMM approach and rule based approach, in this work the HMM model based stochastic tagger is constructed which is based on bigram probabilities then after addition rule based tagger was increase up the accuracy of tagger. In this research work the researchers concluded that the hybrid approach can be used to gain a higher PoS tagging accuracy for Sinhala language.

A Hybrid PoS Tagger for A Relatively Free Word Order language

In the research work of[17] the authors designed hybrid part of speech tagger for Tamil, a relatively free word order, morphologically productive and agglutinative language. In this work the researchers was use both combination of a HMM based statistical PoS tagger and a Rule based PoS tagger. The system works first, the HMM tagger trained using small corpus then given new sentence into tagger and they are tagged. There may be untagged word due to the limitation of algorithm and the amount of training corpus used. Those sentence or words which are not tagged given to rule based system and tagged. The authors used Viterbi algorithm for HMM tagger and to evaluate the accuracy of the taggers by using precision and recall. Hence, after analyzing the result of the tagger, they concluded that the hybrid is good accuracy for the correctness of tagged output.

Hybrid PoS Tagger

In[18] the author designed part of speech tagger for Romanian by using hybrid approaches. The researcher combined statistic model with rule based model. In this work, reducing the tagging ambiguity by PoS dictionary before classifying the input word then tagging input word by using statistical tagger and finally correcting the error of statistical tagging by rule based system was attempted. Finally, the author presented the result of the model for Romanian hybrid tagger which is importantly enhancement of the tagging precision.

Hence, it is observed from the review that many PoS tagging researches were done using rule based, stochastic, CRF, ANN and hybrid approaches for different languages. Some of the research works presented better accuracy using hybrid approaches rather than the individual method. So, in this research, a hybrid approach has been used to tag Wolaita language sentences. The approaches to be integrated for this experiment are transformation based learning and HMM method to develop PoS tagger for Wolaita language. There are around four reviewed researches are related with this research study; However, “PoS tagger for Kafi-noonoo Language” paper is the most related one. Because, the language is categorized on the Omotic family group and it uses Latin script. The model developed for the Kafi-noonoo Language sentence tagger cannot be applied on the Wolaita Language because, they are morphologically different. As the indication of previously done research study on Wolaita Language PoS tagger model though using HMM and CRF method independently; however, the final output produced less accuracy because of the researcher has used small data. But, HMM approach requires large amount of data for better performance. In other previous work researcher developed PoS tagging for Wolaita language using TBL approach and improved the previous work of Berhanu but TBL approach accurate by small corpus. Consequently, this research study follows hybrid approach as a method with relatively large corpus in order to improve the accuracy of tagger.

3. Tags and Tag set of Wolaita Language

For this study 26 tagsets are used from the work of [19] and described in this section.

Table 1: Wolaita Language Tagsets and Description

No	Basic class	Derived class	Description
1	Noun	NN	All common and proper nouns singular and plural tag
2		NP	Noun not separated form preposition tag
3		NC	Noun not separated form conjunction tag
4		NV	Noun with verbal ending
5	Pronoun	PC	Pronoun with conjunction
6		PPRP	Pronoun with preposition tags
7		INTP/WHP	All Interrogative pronoun tags
8		PP	All personal pronoun
9	Verb	VV	All main verb tags
10		VI	All infinite verb tags
11		VC	All subordinate verb including conjunctions and preposition tags
12		VR	All relative verb tags
13	Adverb	ADV	All adverb tags
14		ADVC	Adverb with conjunction or preposition
15	Adjective	ADJ	All adjective tags
16		ADJC	Adjective with conjunction or preposition tags
17	Numerals	AJN	All cardinal numeral tags
18		ON	All ordinal numeral tags
19		CNN	Cardinal with conjunction tags

20	Preposition	PRP	All preposition tags
21		PRPC	Preposition with conjunction tags
22	Conjunction	CJ	All conjunction tags
23	Interjection	INT	All interjection tags
24	Determinant	DD	All determinant tag
25		DPRP	Determinant with preposition tag
26	Punctuation	PUN	All punctuation tag

4. METHODS

A research methodology defines what the activity of research is and how to proceed. In this study the researcher used design science research method to follow to design PoS tagger model for Wolaita language. In this research, the following methods were used to achieve the goal of the research. These are literature review, data collection, data preprocessing, design and implementation and test and evaluation.

4.1. Literature review

In this stage reviewing relevant literatures conducted to gain deep understanding of the study area. To have a better understanding of the problem domain, reading of books that related with this study, journals and research articles important to the research topic was started. In this study, literature review is very important stage to understand the word class and the morphological property of Wolaita language. In this phase to review the literature in the area of PoS tagging based on HMM model and rule based model or transformation based learning (TBL) and the combination of both (hybrid) and discuss with linguistic and expert for more understanding of Wolaita language.

4.2. Data collection

The required data for this study was collected from different sources. These are online bible in Wolaita language, social media in Wolaita language (Wogetta FM 96.6), and also different reference materials of Wolaita language which are provided by the department of Wolaita language in Wolaita Sodo University. From these three data sources the researcher collected 1256 sentences or 15,268 words.

4.3. Data preprocessing

Data preprocessing is an essential phase in any scientific study work. In this stage collected data are preprocessed by removing foreign words[1] and correcting spelling errors before annotating the corpus. After preprocessing the collected data, corpus annotation done manually with help of language experts. Totally, the preprocessed data for this study is 1256 sentences or 15,268 words. After annotation of corpus to preprocess by using system to segment and to tokenize before giving the data for the learning algorithm. The challenging feature in Wolaita language POS tagging is the complexity of the morphological features in WL and in Wolaita language quotation mark is used for two purposes, one purpose is as a quotation mark and the second usage is using as a character to form word . So, this a big challenge when the researcher preprocess the given corpus.

4.4. Design and implementation

Many algorithms have been applied for part of speech tagging including hand-written rules(rule-based tagging), probabilistic methods (HMM tagging and maximum entropy tagging) and Artificial neural network, as well as other methods such as transformation based tagging, memory-based tagging and combination(hybrid) approaches[20]. The research design combines two approaches. The approaches used here are the rule based (Transformation Based Learning (TBL)) and HMM by using Viterbi algorithm. From HMM, rather than using the built-in n-gram function, Viterbi algorithm was used. Rule based part of speech tagger is an approach that solves the problem of assigning the part of speech tags to words in sentences using rules extracted from language experts based on the morphemes attached onto words. These rules can be manually prepared by linguistic professionals and machine learned rules or transformation based learning. Transformation based learning is machine learning technique to generate rule by comparing manually tagged corpus with temporary corpus which is tagged by initial tagger; Transformational based learning takes un annotated corpus as input which goes through the initial state tagger. This initial state tagger assigns a tag that is most likely. This initial

.

tagger produced a temporary corpus as an output then the temporary output corpus by the initial state tagger compared with the goal corpus which was manually tagged and expected to be correct. The corpus passes through the learner iteratively to derive rule for transformations. Each of these derived rules is examined by applying it to the temporary corpus and comparing the result with the goal corpus. Based on this comparison the highest score is applied to the text and is produced as ordered list of rules and added to the result list. The process continues until temporary corpus match with goal corpus or no change in rule.

In this research work, the machine learned rules used rather than using handcrafted rules because which consumes more time and needs skilled linguistic professionals. Machine learned rules can be obtained on the course training of tagger. That means a model is made to automatically learn and store rule called brill Transformation from the training corpus to be provided[12]. Transformation-based learning (TBL) is a rule-based algorithm for automatic tagging of parts-of-speech to the given text[11].Transformation based learning transforms one state to another using change rules to find the appropriate tag for each word.

From the stochastic approach the hidden Markova model (HMM) tagger was designated to tag the words based on the most probable path of the word on a given sentence.

4.4.1. Hidden Markov Models

HMM is the most widely used technique for part of speech tagging in stochastic approach[21]. It is the probabilistic function of Markov process, a process which moves from state to state, left to right on the states, to find optimal state sequence[13]. A Hidden Markov Model (HMM) allows about both observed Model events (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model[10].

Training of the HMM Tagger

This research work implemented HMM tagger by using Viterbi algorithm. This algorithm perform ML to minimize the complexity of HMM tagger in terms of time and memory requirement. The Viterbi algorithm finds the best tag sequence without explicitly computing all sequences.

Design of Wolaita language HMM Tagger

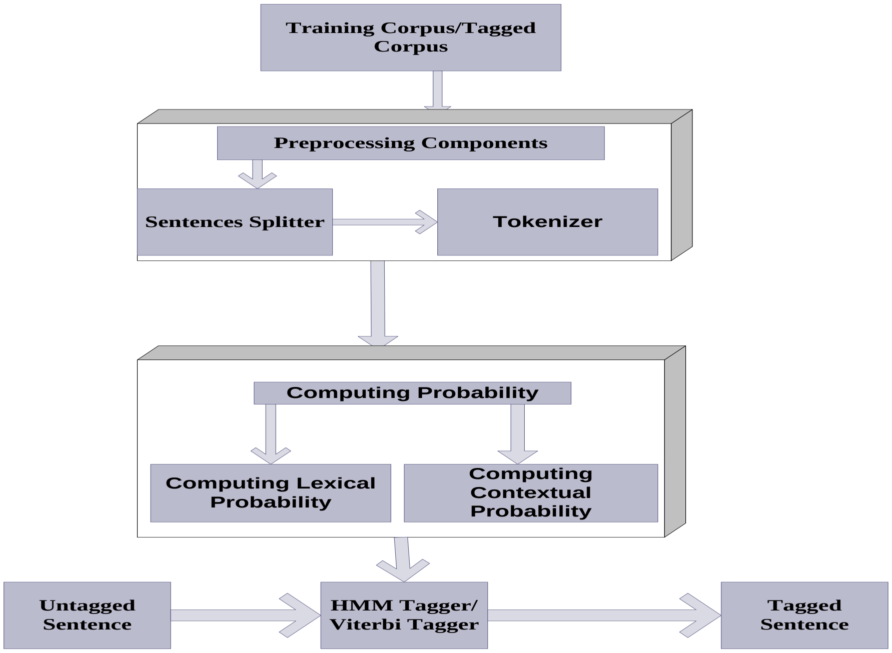

Training Corpus/Tagged Corpus
Preprocessing Components
Sentences Splitter
Tokenizer
Computing Probability
Computing Lexical Probability
Computing Contextual Probability
Untagged Sentence
HMM Tagger/ Viterbi Tagger
Tagged Sentence

Figure 1: Design of HMM Tagger

Description of process in developing model is briefly described below:

i. Tagged corpus

Corpus is most important part of natural language processing (NLP) task like part of speech tagging. Corpus may be in two different forms: untagged and tagged or annotated corpus. For the purpose of this study, the researcher collected data from three domains and preprocessed the collected data. Annotation of corpus is tagging process adding linguistic information to an electronic corpus of written or spoken language data. Tagged corpus is a collection of textual data that contains linguistic information. Whereas untagged corpus is a collection of text without linguistic or grammatical information or untagged words.

i. Preprocessing components

In this study, preprocessing components have three major parts; sentence splitter, tokenizer and tagset analyzer. A processed corpus is input for the model. At first, the content of corpus is given to sentence splitter module to split the text into sentence by using the Wolaita language sentences end markers such as **'.','?'**. The splitted sentence is given to tokenizer module to split the string into words and punctuation. During training phase tagged words were tagged in the form words/tags. The tagset analyzer extracts the tags from the output of the tokenizer. This extracted tagset is used for HMM tagger to tag a new text[22].

ii. Computing probability

Computing lexical and contextual probability from training set; lexical probability from training data set by count the frequency of word with given tag divided by total number of word in training data set. Lexical probability contains the likelihood probability of the word in training corpus and contextual probability contains the transition probability of the word in the training set. These probabilities are given to Viterbi model to pick the optimal path of the word from the two models (lexical and contextual probability). Finally the Viterbi matrix analyzes the most probable path and assign the word with its tags and tag sequence generator generate the word with assigned tags.

iii. HMM Tagger

HMM tagger is the main class of the tagging program/model. It takes untagged sentences as input and produces a tagged sentences or text as output. For instances:

" Neeni tanaara de'iyoogan taani daro ufayittas. "

The tagger take the above sentence as input then to tag words using extracted tag in training time based on the best path of word to produce tagged sentence as output. The following sentences as output of tagger.

“Neeni/PP tanaara/PPRP de’iyoogan/VR taani/PP daro/ADV ufayittais/VV./PUN”

4.4.2. Transformation Based Learning (TBL)

Transformation based learning is rule based system that automatically extracts and learns linguistic information (morphological and contextual information) from correctly annotated corpus without human intervention or expert knowledge. It only requires a sample of correctly manually annotated corpus. TBL is machine based rule learning algorithm that derives lexical and contextual information or linguistic rule from the training corpus and likely part of speech tag for a word. Once the training of the model is completed, the TBL tagger can be used to annotate new untagged Wolaita language sentences based on the tagset of the training corpus. TBL is two stages: initial state tagger and learning phase. In the development of transformation based learning, the TBL tagger take untagged Wolaita sentences as input and initial tagger tag likely tag for the words in the untagged sentences, then result temporary corpus as output. Then the second stage (learning phase) take two corpus which is temporary corpus and goal corpus which is manually tagged corpus expected as correctly tagged corpus then the learning phase compared temporary corpus with goal corpus for rule derivation. The temporary corpus passes through the second stage which is learning phase iteratively to derive rule transformation. The learning stage learns continues until no change of rule to improve temporary corpus compared with reference or goal corpus. In the development of this, the learning stage produce ordered list of rules which can be applied and tagged untagged sentences.

4.4.3 Hybrid Tagger

In this research, the implemented hybrid tagger is two-step process. The first step process is performed by HMM tagger and second step process is performed by rule based tagger. The HMM tagger first take untagged text and assigns tag to raw text based on probability and proved optimal level of tag sequence. The threshold value is not attained in HMM tagger, it is corrected by second step process which is rule based tagger. Rule based tagger correct the error of tag sequence by applying transformation rule that it has learned during training time. The threshold value is fixed based on the result of performances of hybrid tagger in experiments.

The HMM tagger use Viterbi algorithm to find the optimal probabilities of the tag/word pairs and the probabilities of optimal path. Thus, it is possible to compare the probabilities of each word tag is greater than the fixed threshold value. When the probability of the assigned tag by HMM tagger for the given word is greater than the fixed threshold value it ensure that the assigned tag by HMM tagger is correct tag or does not need correction by rule based tagger. Otherwise, the word is given to rule based tagger to correct tag. hence, the proposed hybrid tagger accept untagged text and tag using HMM tagger which acts as an initial tagger for the raw text to be tagged and the rule based tagger correct the output of HMM tagger by applying rules if the predetermined threshold value is not attain.

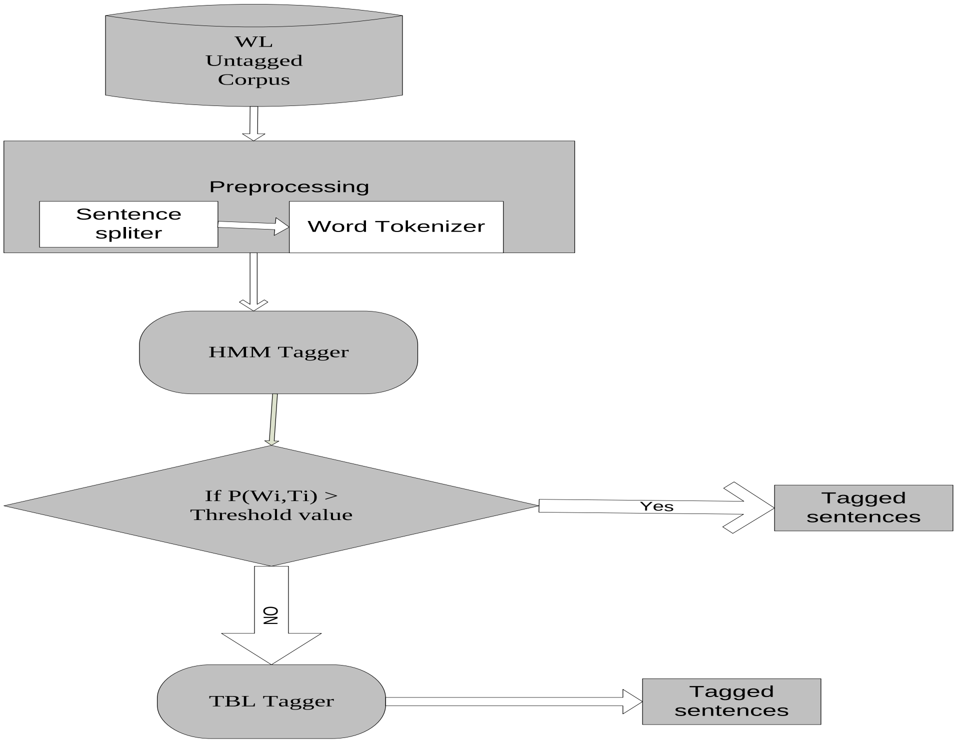


Figure 2: Design of Hybrid Tagger

The untagged texts of Wolaita language are given to the preprocessing component. Then the word sequence Wi is given to the HMM tagger as an input; the HMM tagger assign the tag sequence Ti using Viterbi algorithm (which find the best optimal path) and the output of the HMM tagger is the word tag sequence (Wi,Ti). This word tag sequence is given to the output analyzer that checks whether the determined threshold value for a word Wi is achieved or not. The threshold value is a value used for checking the sureness level of tagging a given sequence of words. So, the output analyzer decides based on the threshold value. Consequently if the threshold value of a word tag pair is lower than the fixed threshold value, a fixed window size, in this case a window size of two which implies bigram of words is given to the transformation based learning tagger for correction and the transformation based learning tagger produce the corrected tagged words. Otherwise the HMM tagger output is the final output of the word to be tagged. This process is repeated until the HMM tagger tagged words are checked by threshold value and corrected by transformation based learning tagger.

4.5. Test and evaluation methods

The author used percentage split method to test and evaluate the model. To measure the accuracy, the manually entire corpus was divided into training data set and test data set. The gold standard tags are used to compare and evaluate percentage accuracy of the model. The system tags test sentences based on the trained knowledge and then estimate model by count correct tags from test sentences and divided by the total number of test sentences (incorrect and correct tags).

$$\%\ \text{accuracy} = \frac{\text{Number of correct tags}}{\text{Number of incorrect tags} + \text{number of correct tags}}$$

5. RESULT and DISCUSSION

This section describes the experimental results and respective discussion for application of hybrid approach for Wolaita language part of speech tagger. In order to implement this study the researchers have used python programming language packages. To develop the HMM model, Transformation based learning and Hybrid (HMM & Transformation based learning) model for Wolaita PoS tagger. After comparing the two models performance, the author used HMM as initial tagger and TBL as a corrector tagger.

5.1 Experiment Results

In this research, the author evaluated three experiments with similar validation method and with the same corpus. These are HMM tagger, TBL and hybrid tagger. In hybrid tagger there are five experiments by using different threshold value. The researcher used percentage split evaluation method in each individual experiment. As said by this percentage split evaluation method, the whole corpus is split into training dataset and testing dataset. The whole corpus used for this research is 1256 sentences or 15,268 words. To choose the splitting size of training and testing dataset the following experiments were done by using different percentage split.

Table 2: Experiment Results of HMM Tagger using different Train /Test Split

	70/30(70% training set and 30% test set)	80/20(80% training set and 20% test set)	90/10(90% training set and 10 test set)
HMM Tagger Accuracy	72.52%	73.91%	88.14%

In HMM tagger there is three experiments to decide the percentage of training set data and test data. The researcher used three percentage split: those are 70/30, 80/20 and 90/10. The accuracy of HMM tagger is 72.52%, 73.91% and 88.14% of 70/30, 80/20 and 90/10 respectively. From those 90/10 split accuracy of tagger is higher. So, the researcher decides 90/10 percentage split based on the experiment result. Table 10 shows experiment result of transformation based learning tagger for each percentage split.

Table 3: Experiment Results of TBL Tagger using different Percentage Split

Initial Tagger	70/30(70% training set and 30% test set)	80/20(80% training set and 20% test set)	90/10(90% training set and 10% test set)
Unigram Tagger	77.51%	79.80%	89.59%
Bigram Tagger	75.89%	78.48%	90.99%
Trigram Tagger	77.19%	79.99%	93.19%

The above table 3 showed that the experiment result of TBL tagger of three different percentage split and the percentage split of 90/10 (90% training data set and 10% test data set) is outperformed others. So, the researcher decided 90/10 percentage split for this study based on the experiment result.

5.1.1. Experiment Result for HMM Tagger

In order to test the performance of the HMM tagger like that of transformation based learning tagger. From the entire corpus, 90% is used for training and the remaining 10% is used for testing purpose. To conduct the experiments, first the entire training dataset divided into ten parts, started training HMM tagger the first 10% sentences of training dataset. After the HMM tagger is trained using 10% of training dataset, performance of trained tagger is measured by test dataset and repeated this process by adding training dataset by 10% until 100% training dataset.

Table 4: Shows different Experiments with different Portion of Training-dataset for HMM Tagger

Training-set	10%	20%	30	40%	50%	60%	70%	80%	90%	100%
HMM-tagger-accuracy	76.45%	77.61%	77.32%	77.26%	78.08%	78.43%	78.37%	78.89%	79.01%	88.14%

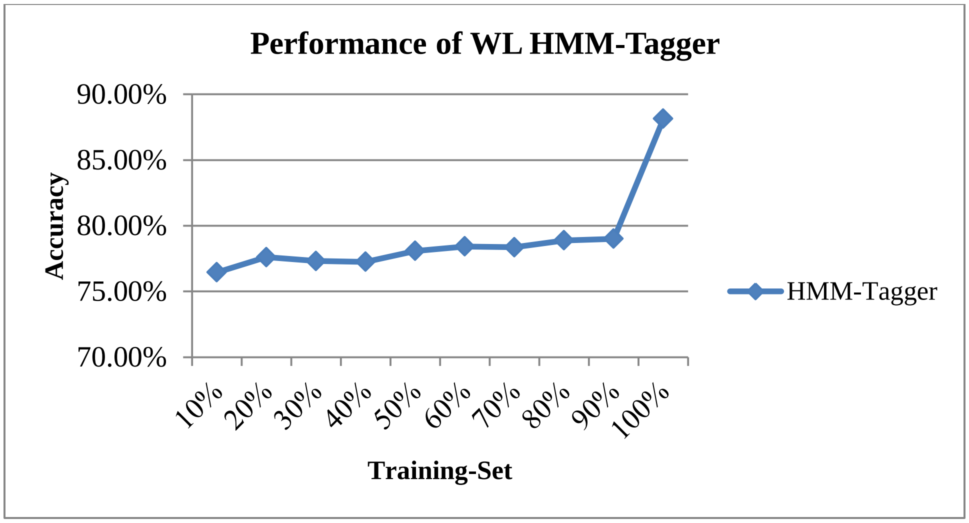


Figure 3: Performances of HMM Tagger (Viterbi)

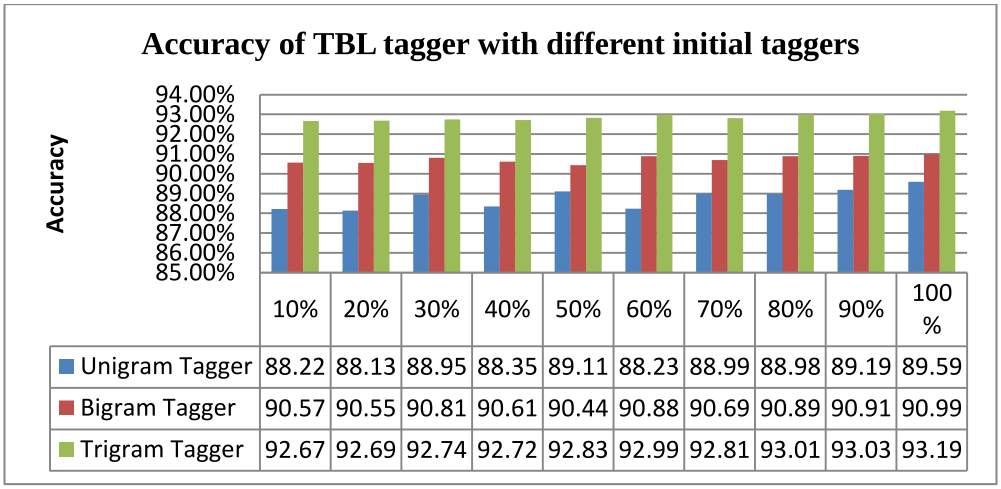
The above Figure 3 shows the performance of HMM tagger for Wolaita language when the system takes different amount of training dataset incrementally and the same amount of test dataset. Also it describes that the characteristics of HMM tagger, because the HMM tagger require a large amount of training dataset to give better performance. In this case, when training dataset is 100% the accuracy is 88.14%.

5.1.2. Experiment Result for Transformation based learning

The performance of rule based tagger tested using ten different experiments with different portion of training set by three different initial taggers namely unigram, bigram and trigram tagger.

Table 5: Shows different Experiments with different Portion of Training dataset and Initial Taggers for TBL Tagger

Training set	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Unigram Tagger	88.22 %	88.13 %	88.95 %	88.35 %	89.11 %	88.23 %	88.99 %	88.98 %	89.19 %	89.59 %
Bigram tagger	90.57 %	90.55 %	90.81 %	90.61 %	90.44 %	90.88 %	90.69 %	90.89 %	90.91 %	90.99 %
Trigram tagger	92.67 %	92.69 %	92.74 %	92.72 %	92.83 %	92.99 %	92.81 %	93.01 %	93.03 %	93.19 %



	10%	20%	30%	40%	50%	60%	70%	80%	90%	100 %
🟦 Unigram Tagger	88.22	88.13	88.95	88.35	89.11	88.23	88.99	88.98	89.19	89.59
🟥 Bigram Tagger	90.57	90.55	90.81	90.61	90.44	90.88	90.69	90.89	90.91	90.99
🟩 Trigram Tagger	92.67	92.69	92.74	92.72	92.83	92.99	92.81	93.01	93.03	93.19

Figure 4: Rule Based Tagger Performance for Wolaita Language

The above figure 4 show the experiment results of transformation based tagger. In this study, the researcher used three initial taggers (unigram, bigram and trigram tagger) with different portion of training dataset. The model training first stared by 10% of training dataset and test dataset to test the accuracy of TBL tagger and continue training the model by add 10% of training dataset until 100% training set. When the training dataset increase the accuracy of the model increase; the performance of TBL tagger is 89.59%, 90.99% and 93.19% of unigram, bigram and trigram respectively. However, the performance of transformation based learning or rule based tagger used trigram taggers as initial tagger greater than bigram and unigram tagger.

5.2.3. Experiment Result for Hybrid Tagger

The designed hybrid taggers for Wolaita language are combined HMM and rule-based tagger. In this hybrid model, the HMM tagger first annotate the word sequence within the sentence and if the desired threshold value of the given sentence is not attained, the sequence of word is given to the rule based tagger for correction. In this research, the HMM tagger is used as initial annotator and rule-based tagger as a corrector. The fixed threshold value for this study is 0.6 since taking threshold value less than 0.6 does not bring significant difference on the performance of the tagger which means less than 0.6 give to TBL tagger for correction. Rule based tagger corrects more words when the threshold value increase. As a result of this, the hybrid tagger gives better performance when the threshold value goes up, As it is presented in table 6. Hence, the fixed threshold value to 0.6, performance of 94.82% is obtained. But the threshold value increase to 1, the performance of tagger less accurate because the probability always between zero and one. Table 6 indicates the performance of hybrid tagger with different threshold value.

Table 6: Performance of hybrid Tagger

Threshold-Value	0.1	0.2	0.4	0.5	0.6	0.8	1
Performances	90.696	92.012	93.932	94.155	94.826	94.822	90.400

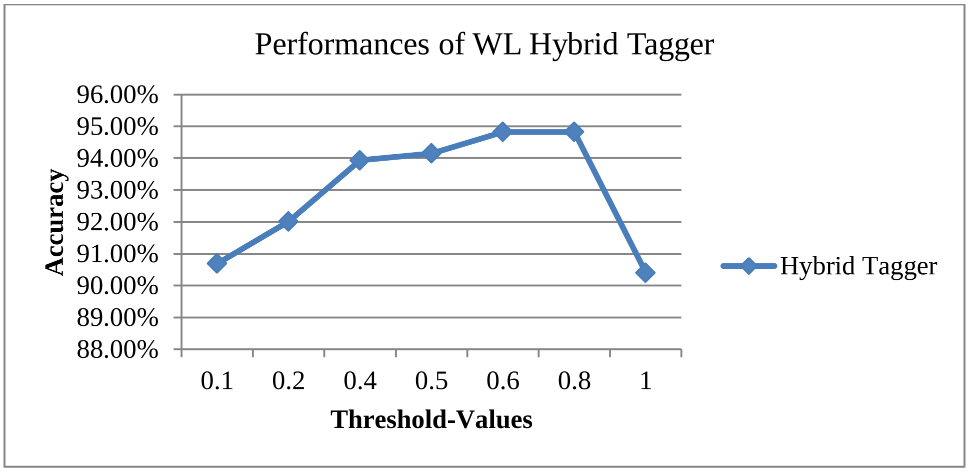


Figure 5: Performance of Hybrid Tagger

The above figure 5 shows the performance of hybrid tagger that consists HMM and TBL tagger. In the graph the researcher used different threshold value between 0 and 1. The performance result of hybrid tagger in threshold value 0.6 and 0.8 is 94.826, 94.822 respectively. The fixed threshold value is 0.6 based on the performance result because the performance result of threshold value 0.8 is relatively less accurate than the threshold value of 0.6. The threshold value increase, transformation based learning tagger corrects more words. As a result of this, the hybrid tagger gives better performance as the threshold values increase. But when the threshold value closed to 1.0 the accuracy goes down because most of the time the probability is not greater than or equal to 1, in this case TBL tagger tags all the words. So, the fixed threshold value 0.6, overall performance of hybrid tagger 94.82% is achieved.

6.1 Conclusions

Part of speech tagger is one of the ground level application areas of natural language processing. PoS tagging is the process of classifying corresponding PoS tag for a word in sentence. Different researchers developed PoS tagger for different languages by using different approaches. For this study, a hybrid approach used that combines HMM and TBL tagger at sentence level is designed for Wolaita language. This paper describes a sequence of different PoS tagging experiments and to develop this hybrid model the HMM and TBL tagger were developed and evaluated three tagger individually those are rule based tagger, HMM tagger and hybrid tagger.

For the purpose PoS tagger development 1256 sentences (15268 words) and 26 tagsets are used. The corpus and tagsets are prepared manually for this research development; because the language there is no standard corpus and tagsets for natural language processing. The whole corpus is divided into training and testing dataset. The author decided to use 90% training dataset and 10% testing dataset of entire corpus based on experiments of above table 2 and 3. For implementation and experiment of PoS tagger NLTK and Python version 3.5.0 are used. To evaluated the performance of three types of tagger namely HMM tagger, rule-based and hybrid tagger with the same training dataset and testing dataset conducted by different experiments. As a result, 88.14%, 93.19% and 94.82% performances are obtained for HMM, rule-based with trigram initial state tagger and hybrid taggers respectively.

The performance of hybrid tagger was tested on different threshold values and threshold value of 0.6 scored better performance than other threshold values. So, 0.6 was used as fixed threshold values of Wolaita language hybrid tagger. When compare the accuracy of hybrid tagger with HMM and rule-based tagger individually, hybrid tagger increased by 6.68 than HMM tagger and 1.63% than rule-based tagger.

Among these three tagger the hybrid approach outperformed the rule based and HMM tagger. Thus, the hybrid approach is better for classifying tag for word of Wolaita language at the sentence level. So, based on their performance, the study has concluded that the hybrid tagger performs better than two approaches HMM and rule-based tagger taken separately.

Reference

[1] Mamo Getachew, 2009 “Part-of-Speech Tagging for Afaan Oromo Language,” Addis Ababa Univerity,.

[2] Perkins J., 2014, *Python 3 Text Processing with NLTK 3 Cookbook*. PACKT Publishing Ltd.,.

[3] Bade G. Y. and Seid H., 2018 ,“Development of Longest-Match Based Stemmer for Texts of Wolaita Language,” vol. 4, no. 3, pp. 79–83,.

[4] Abebe Tewodros Gebreselassie, 2009 ,“Text-To Speech_Synthesizer_For_Wolaytta,” Addis Ababa University,Ethiopia,.

[5] Fanta Habtamu. , 2010 “Speaker Dependent Speech Recognition For Wolaita Language,” Addis Abeba University, Msc Thesis.

[6] Lessa L., 2003, “Development of Stemming Algorithm for Wolaita Language,” Addis Ababa University, Msc Thesis.

[7] Mamo Getachew, 2005, “Automatic part-of-speech tagging for Amharic language an experiment using stochastic Hidden Markov Approach,” Addis Ababa University, Msc Theesis,.

[8] Biadgo Y., “Application of multilayer perception neural network for tagging part-of-speech for Amharic language,” Addis Ababa University, Msc Thesis.

[9] B. H. Ganta, 2015 “Part Of Speech Tagging for Wolaita Language,” Addis Ababa University, Msc Thesis,.

[10] G. Emiru, 2016 “Development of Part of Speech Tagger using Hybrid Approach,” Addis Abeba university, Msc Thesis,.

[11] Mekuria Zelalem, 2013 “Design and Development of Part-of-speech Tagger for Kafi-noonoo Language,” Addis Ababa University, ,Msc Thesis,.

[12] A. L. and M. M. Meryeme Hadni, Said Alaoui Ouatik, 2013, “Hybrid Part of Speech Tagger for Non-Vocalized Arabic Text,” *Int. J. Nat. Lang. Comput.*, vol. Vol.2.

[13] Abreha T. G., 2010 ,“Part of Speech Tagging for tigirgna Language,” Addis Ababa University,Msc Thesis.

[14] K. Mohnot, N. Bansal, S. P. Singh, and A. Kumar, 2014, “Hybrid approach for Part of Speech Tagger for Hindi language,” vol. 4, no. 1,

[15] K. N. R. N. Francis Merin, 2014 ,“Hybrid Part of Speech Tagger for Malayalam,”

IEEE,.

[16] A. R. W. Dilmi Gunasekara, W.V.Welgama, 2016 ,“Haybrid part of Speech Tagger for Sinhala Language,” in *International Conference on Advances in ICT for Emerging Regions(ICTer)*,.

[17] P. Arulmozhi, “A Hybrid POS Tagger for a Relatively Free Word Order Language,” *ACADEMIA*.

[18] R. Simionescu, 2011 ,“Hybrid Part of Speech Tagger,” in *Proceedings of the Workshop “Language Resources and Tools with Industrial Applications,”*.

[19] Shirko Birhanesh F., 2020, “Part of Speech Tagging for Wolaita Language using Transformation Based Learning (TBL) Approach,” *IJESC*, vol. 10, no. 9, pp. 27214–27222,.

[20] J. H. M. Jurafsky, Daniel, *Speech and Language Processing: An Introduction to Natural Language Processing, omputational Linguostics, and Spech Recognition*. PEARSON Prentice Hall,New Jersey, 2006.

[21] S. D. S. S. A. Basu, “A Hybird Model for part of speech tagging and its application to Bengali,” vol. 1, 2004.

[22] E. L. Steven Bird, Ewan Klein, “Natural Language toolkit and Python,” First Edit., 1005, Gravenstein Highway North, Sebastop: O’Reilly Media,Inc, 2009.