Implementation of Hidden Markov Model on Lagos Nigeria Women Union (NWU) Yoruba Speech Corpus

A.O. Enikuomehin and A.A. Tijani *

Department of Computer Science, Lagos State University, Lagos Nigeria.

E-mail: toyinenikuomehin@gmail.com*

ABSTRACT

Part of Speech (POS) tagging is the process of assigning correct corresponding grammatical category or part-of-speech label that best suits the definition of a word as well as its context in a particular position of a sentence in which it is used. This is usually a foundation stage for many Natural Language Processing (NLP) applications. Yoruba language is one of the most spoken languages in West Africa but has the least Part of Speech (POS) tagging research which results in very few NLP tools such as corpus, tagset, etc. Therefore, this paper implements a stochastic learning approach algorithm, Hidden Markov Model (HMM), on Yoruba corpus. Five hundred manually annotated sentences from the main corpus are used as training data and 100 sentences distinctive of the training data are used as test data.

(Keywords: Yoruba, corpus, POS Tagger, part of speech, NLP, natural language processing, speech synthesis, speech recognition, information retrieval).

INTRODUCTION

Part of speech tagging (POS tagging) is key in Natural Language Processing (NLP) aspects such as Speech Synthesis, Information Retrieval, Speech Recognition and machine translation. It can also be referred to as word-category disambiguation or grammatical tagging which is a process of assigning part of speech (label) to individual word in a document according to its contextual meaning and definition (Francis, 2014).

This usually has a lager tagset which includes further classification of these basic grammar classes (e.g., representing a plural noun by NNS (Noun, non-singular)). In this work, we manually annotated Yoruba corpus (Lagos NWU Yoruba speech) to train a Hidden Markov Model algorithm. This algorithm uses statistical methods to calculate the most probable tag sequence for sequence of text. The trained model is evaluated using the Tag-wise precision recall method.

Yoruba Language

Yoruba is the third most spoken language in West Africa with over fifty Million speakers. It is one of the three major languages in Nigeria with the belief of Oduduwa (son of Olodumare) as the ancestral speaker. This language first appeared in writing in the 19th century with the first publication produced by John Raban (Adedjouma, Aoga, and Igue, 2013). Yoruba language has twenty five (25) distinct alphabets: twelve (7 oral and 5 nasal) vowels; eighteen consonants. It's a toner language with three tones: high, low, and middle. The various Yorùbá dialects in Nigeria can be classified into three major dialect areas viz:

i. North-West Yorùbá NW which include Ibadan, Oyo, Ogun and Lagos (Éko areas),

ii. Central Yoruba(CY) which include Igbomina, Yagba, Ife, Ekiti, Akure and Ijebu areas, and

iii. South- East Yoruba (SEY) which include Okitipupa, Ondo, Owo, Sagamu and some parts of Ijebu (Abiola, Adetunmbi and Oguntimilehin, 2015).

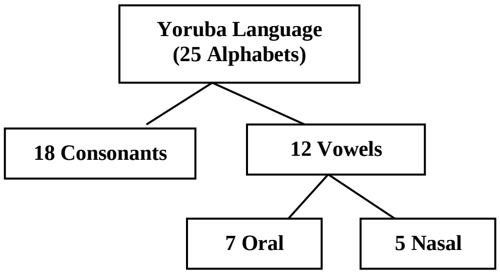


Figure 1: Yoruba Language Alphabets.

Morphology of Yoruba Language

Morphology in relation to linguistics simply means the study of the forms and structures of word in a particular language. Morphological analysis is a fundamental piece of NLP; it is pivotal to language comprehension and computerization as it represents word development in dialects.

Yoruba language is semi–agglutinative language which implies a language where words are comprised of linearly successive morphemes with components representing a meaningful morpheme. As a rule, the word order is Subject-Object-Action-word (Verb), though other word order applies since the language is verse (Enikuomehin, 2015).

The language order can be:

1. Sentence = Verb + Subject (Noun/Pronoun) + Object(Preposition/Noun)

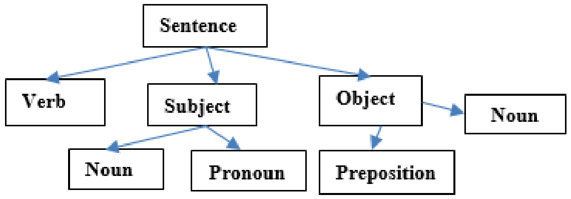


Figure 2: Tree Diagram for Verb-Subject-Object (VSO).

2. Sentence = Subject + Verb(Noun/Pronoun) + Object(Adjective/Adverb)

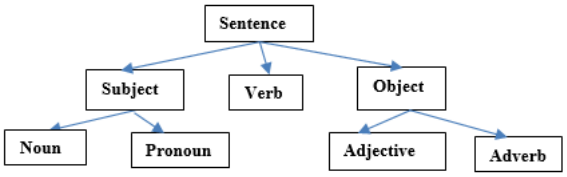


Figure 3: Tree diagram for Subject-Verb-Object (SVO)

Tagset

Words from any existing language in the World can take the form of a noun, pronoun, verb, adverb, adjective, preposition or conjunction. Part of Speech taggers is saddled with the role of tagging or labelling words with appropriate tags. The set of tags from which taggers chooses the appropriate option for a particular word is known as Tagset. Therefore before the conclusion of a tagset, the following properties should be considered:

i. Fineness/coarseness: the fineness of the tagging should be known i.e. whether the tags will allow for precise distinction of the various features of POS of the language e.g. plurality, gender etc. or whether the tagger would only provide the different lexical categories.

ii. Syntactic function/lexical category: The lexical category of a word can be different than the POS of the word in a sentence, and the tag set should be able to represent both.

iii. New tags/tags from a standard tagger: It has to be decided whether an existing tag set should be used, or a new tag set should be applied according to the specifics of the language on which the tagger will work (Hasan, 2006).

Problem Statement

The study of POS tagging algorithms on Yoruba language has not been adequately explored. There are a few Corpora of Yoruba Language that are publicly available and the few available are motivated by specific research agenda. Also, there is hardly any freely accessible annotated corpus, Tagset or other NLP tools on the internet for Yoruba language; most works done on Part of Speech Tagging algorithms are on English language and other indigenous languages (Arabic, Bahasa-Indonesia, Bangla, Malay, etc.) rather than Yoruba language. So, we implemented Hidden Markov Model algorithm on Yoruba Corpus for analysis.

History of Taggers

The pioneer tagger (TAGGIT) was used in 1971 for the initial tagging of the Brown's Corpus (BC), since which a lot of effort has been committed to upgrade the quality of tagging process in respect to efficiency and accuracy. TAGGIT which was a rule-based technique was developed by Greene and Rubin in 1970 which achieved an accuracy of 77%. Eric Brill's tagger achieved an accuracy of 96% in 1992 which was later improved to 97.5 in 1994 (Khoja, 2001).

POS tagging models usually consist of unigram, bi-grams, and tri-grams (sequences of one, two or three consecutive tags, respectively). Once the n-gram probabilities have been estimated, new examples can be tagged by selecting the tag sequence with highest probability. This is roughly the technique followed by the widespread Hidden Markov Model taggers (e.g., Claws).

Claws, a probabilistic version of TAGGIT uses bigram model and was developed in Lancaster University. It achieved an accuracy of 97% while Church's PARTS tagger uses trigram model and was developed in1988. The Xerox tagger which was developed by Doug Cutting achieved an accuracy of 96% in 1992. These statistical models involve some kind of learning, supervised or unsupervised, of the parameters of the model from a training corpus (Padr, et al., 2000).

POS Tagging Approaches

a. **Supervised Approach:** this is either ruled based or stochastic approach.

i. **Rule Based:** This model uses the application of contextual meaning and set of hand written rules also referred to as context frame rules to assign POS tag to a given set of text (corpus) (i.e., tagging is done in two stages, the first stage uses the dictionary to assign list of likely part of speech while the second stage uses set of context frame rule to narrow the list to one part of speech) (Francis, 2014).The rules are based on knowledge of the specific language which may consist of a large number of morphological, lexical and syntactical information. These rules can be obtained manually, they are handcrafted by linguistic professionals. The manual way of getting rules is tedious and time taking since it requires linguistic professionals to set rules. Moreover, it is inconsistent and subjective as it is determined by the understanding of one or more linguistic specialists and their skill and knowledge of the specific language.

ii. **Stochastic Approach:** This model uses calculated frequency and statistics or probability of each word in a text (corpus) (i.e., it uses the frequently used tag for a word to tag it anytime it comes across it). (e.g., Viterbi algorithm).

b. **Unsupervised (Bootstrap)**

The use of untagged text (corpus) for training and produce a tagset through induction (P.J. Antony, 2011). This do not require pre-tagged corpora, it rather uses advanced computational methods such as the Baum-Welch algorithm so as to automatically induce tag sets, transformation rules etc. (Kumawat and Jain, 2015).

i. **Neural:** Neural networks consist of large number of simple processing units which are highly interconnected by directed weighted links. This implies that the information processing of Neural Networks (NN) was known in biological nervous system before actually applying it in information processing using computers applications. Neural Networks learn from example by configuring for a specific application such as pattern recognition or data classification. The NN can learn by adapting different behavior on the basis of the data that is given to the network. It is possible to call the NN learning an adaptive learning as the network is able to find properties from the presented data. It is not necessary to tell the network how to react to each data input separately like the conventional programming.

DISCUSSION ON LITERATURE

Many POS tagging techniques have been implemented on English language and many other western languages with a satisfying performance of 96+%. A morphological analyzer indeed provides some POS tag information, but a POS-tagger needs to operate on a large set of fine-grained tags. For example, English language consists of 87 distinct tags, and Penn Treebank’s tag set consists of 48 tags (Chowdhury et al., 2004).

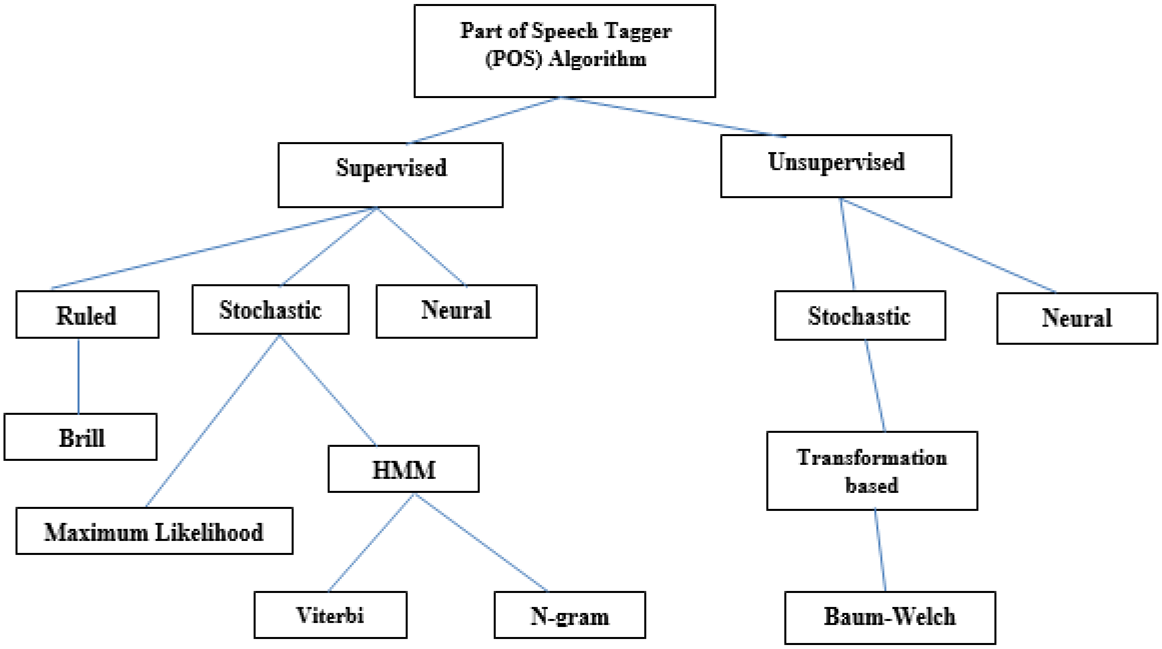


Figure 4: Classification of POS Tagging Algorithms.

Alfan and Ayu in 2010 defined Hidden Markov Model as an established probabilistic method for automatic POS tagger. They stated that many languages that have adapted the HMM method in building automatic POS tagger proved to have better running time than any other probabilistic methods. They then developed a Hidden Markov based POS tagger which uses affix tree to predict emission probability vector for OOV words and utilizing information from dictionary lexicon and succeeding POS tag. The developed tagger uses both first order (bigram) and second order (trigram) Hidden Markov Model (Alfan and Ayu, 2010).

Mohamed, Omar and Ab Aziz in 2011 when working on Malay POS tagging used Trigram Hidden Markov Model (THMM) with the aid of prefix and suffix characters as the guesser for unknown words to achieve 67.9% accuracy on one thousand eight hundred and forty token used for the test(Mohamed, Omar, and Ab Aziz, 2011).

Also, an open source tagger based on Hidden Markov Model (relax) was developed by Marco in 2012, which yielded a result close to state-of-the-art tagger. The tool used to create this part-of-speech tagger is FreeLing which is an open source text processing tool offering a number of language analysis services, such as morphosyntactic tagging, named entity recognition, dependency parsing or sense annotation. FreeLing provides an application programming interface (API) that can be used to integrate language analyses into a more complex processing (Marco, 2012).

Amrullah, Hartanto, and Mustika in 2017 when working on Tagging of Bahasa Indonesia text compared unigram, Hidden Markov Model and Brills tagger. Using Natural Language Toolkit (NLTK), accuracy of the result was calculated using number of correctly tagged token compared to the tagged corpus which yielded result of 88.37% when unigram was used. (Amrullah, Hartanto, and Mustika, 2017).

METHODOLOGY

The implementation of Hidden Markov Model (HMM) algorithm on Yoruba Corpus was done in three stages; Tagset preparation/Corpus preparation, Tagging, and evaluation.

Choice of Corpus/Tagset Preparation

Tagset Preparation: We considered Penn tree tagset as a reference point for our tagset design diverging where necessary. The Penn Tree bank tagging guidelines for English proposed a set of 36 tags, which is believed to be one of the standard tagset for English. However, the number and types of tags needed for POS tagging differ from language to language. In the context of Yoruba language, we did not know of many works on tagset design when we started the work. Thus, a Tagset of 14 distinctive tags was coined out of the pen-tree Tagset for the purpose of this study.

Table 1: Tagset used for Yoruba Language Tagging.

Tag	Description	Tag	Description
NN	Noun, singular or Mass	RB	Adverb
NNP	Noun, plural	DT	Determinant
NNPS	Proper Noun singular	TO	To
JJ	Adjective	VB	Verb
CC	Coordinating Conjunction	CD	Cardinal number
PRP	Pronoun	MD	Modal Verb
FW	Foreign word	IN	Preposition

Corpus Preparation: Annotated Corpus is used in many NLP applications such as POS tagger training and testing, parsing, sentiment analysis etc. In this research work, the annotated corpus used is considered to be a text tagged with the corresponding part of speech tags. In fact, the tagged text i.e. the annotated corpus is thought to represent significant domains of Yoruba language. LAGOS-NWU (Lagos Nigeria Women Union) Yoruba Speech Corpus was our choice of digital resources due to its arrangement and accuracy in words spelling. This is an open source material which may be used free of licensing charges. The texts form this corpus required no cleaning; it has an arrangement of one sentence per line which made it suitable for the study.

LAGOS-NWU Yoruba Speech has 21,728 words containing around 8000 distinct words. This may not be as large as some English corpus which may contain millions of words but it is the best we could get as at the time of this research. 500 randomly selected sentences where manually tagged as there was no readily available tagged corpus of Yoruba language available. The manual tagging is done by the researcher with the aid of English-to-Yoruba Yoruba-to-English Dictionary. Due to the cumbersome nature of manual tagging and time constraint, approximately five hundred (500) randomly selected sentences were tagged from the main corpus. These tagged sentences were used for training of the model (HMM).

Training of the Model: A training corpus consisting of manually annotated sentences (i.e. 500 sentences approximately 2,000 words) from the main corpus was prepared; this is used in the training of HMM model. Adopting the supervised learning approach, the manually tagged corpus served as input which allows the model to learn the rules of the language. The corpus reader reads the contents of the corpus and passes it to the Tokenizer which breaks down the sentence to word level (tokens) using space character.

The tagset analyzer extracts the tags from the words and stores them in the database. The decoding algorithm (Viterbi algorithm) computes the Part of Speech (POS) tag probabilities which are important for finding the sequence of words in the input sentence. These tagged texts were used for training the POS tagger; the trained tagger takes untagged text as an input and tags the words based on the knowledge that it has acquired during the training to produce tagged text as output. Randomly 100 sentences distinctive of the training data were selected for testing the model. This process can be represented using the diagram below:

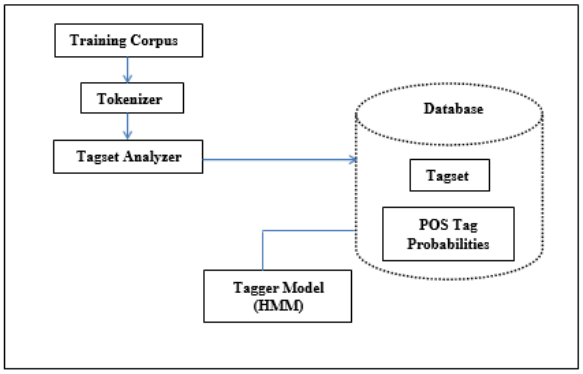


Figure 5: Training HMM.

Decoding Hidden Markov Model

Viterbi algorithm is a dynamic programming algorithm that is usually used for decoding HMM algorithm. It finds the optimal path in the tagging process and reduces the complexity of the HMM core issues ranging from finding the best part of speech tag sequence for a given sequence of words in the input sentence to the polynomial time. For any input sentence of length n, there are $|K|^n$ possible tag sequences. The exponential growth with respect to the length n means that for any reasonable length sentence, this search will not be tractable.

Summarily, the Viterbi algorithm computes the probability of all conceivable paths of the word tag pairs in the input sentence and selects the path of the word tag pair that has the highest probability to be the best path. The probability of a label sequence given a set of observations is defined in terms of the transition probability and the emission probability which is mathematically represented as:

$$p(x_1 \dots x_n, y_1 \dots y_{n+1}) = \prod_{1=1}^{n+1} q(y_i | y_{1-2}, y_{i-1}) \prod_{i=1}^{n} e(x_i | y_i)$$

$$r(y_1 \dots y_k) = \prod_{i=1}^{k} q(y_i | y_{i-2}, y_{i-1}) \prod_{i=1}^{k} e(x_i | y_i) \qquad \textbf{(1)}$$

Such that 'r' is considering the first k terms where $k \in \{1..n\}$ and for any label sequence $y_1 \ldots y_k$. Also, is set $S(k, u, v)$ which is simply the set of all labeled sequences of length k that ends with the bigram(u, v).

We have set of sequences $y_1 \dots y_k$ such that $y_{k-1} = u, y_k = v, \pi(k, u, v)$ which is simply the sequence with the maximum probability and can finally be defined as:

$$\pi(k, u, v) = \max_{\langle y_1 \dots y_k \rangle \in S(k,u,v)} r(y_1 \dots y_k) \qquad \textbf{(2)}$$

$$\pi(k, u, v) = \left(\pi(k-1, w, u) \times q(v | w, u) \times e(x_k | v)\right) \qquad \textbf{(3)}$$

Finally, we have

$$\max_{(y_1 \dots y_{n+1})} p(x_1 \dots x_n, y_1 \dots y_{n+1}) = \max_{u \in k, \in v \in k} \left(\pi(n.u, v) \times q(STOP | u, v)\right) \qquad \textbf{(4)}$$

This algorithm has execution time of $O(n|K|^3)$, hence it is linear in the length of the sequence, and cubic in the number of tags.

Tagging

The choice of Hidden Markov Model, a statistical algorithm for the POS tagging is to ensure robustness of the system. This model is trained to learn the language using annotated sentences which are provided and later used to precisely predict the category of a new word of the language.

HMM is a statistical learning algorithm based on finding the probability (p) of a word along its tag from a given text which can be represented mathematically as:

$$p(W_i, T_i | < S >)$$

Where W_i, T_i are the i^{th} word and the i^{th} tag are in the input Corpus. During the POS tagging, HMM simply finds the most frequent tag from the training corpus. This is done by counting the occurrence of the specific word W_i associated with a tag T_i and dividing it by the total occurrence of the word W_i in the training corpus which can be represented mathematically as:

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

During the training phase, the probability of each word attached to a POS tag is calculated and used during the tagging process of new sentences. It allocates the most probable tag to each word.

The Tokenizer receives untagged text as part of the pre-processing stage for the tagging process. Afterwards, tag sequence is generated for these tokens by the model and this is displayed as the output. The process of finding the optimal sequence of part of speech tags for the given tokens in the input sentence to be tagged is done by Decoding algorithm.

IMPLEMENTATION AND RESULTS

The process of tagging Yoruba Language is implemented using the Natural Language Toolkit (NLTK) and Python as the programming Language.

Python and Libraries

Python is chosen as the programming language for this study due to the following reasons:

i. Python is a simple to adapt yet incredible programming language particularly for text processing in NLP applications. It has efficient high-level data structures and a simple but effective approach to object-oriented programming.

ii. Also, Python exquisite syntax and dynamic typing with its interpreted nature make it a perfect language for scripting and fast application development in different areas especially in NLP on many platforms.

Since Python is used, pyCharm by jetbrain is adopted as the IDE. PyCharm is an intelligent tool for thousands of professional Python developer around the world. It provides programmers with intelligent code completion, on-the-fly error checking and quick fixes, easy project navigation and lots more. Also, other Python libraries utilized for the purpose of this study may include

- Scikit-learn**,** a Python module for machine learning which is built on SciPy. It runs on Python version 3.5 or higher; NumPy version 1.11.0 or higher; SciPy version 0.17.0 or higher.

- Python NumPy which is often referred to as python alternative to MATLAB and often used with SciPy module.

Natural Language Tool Kit (NLTK)

The basis behind the decision to use NLTK tools is that:

i. They are most appropriate tools used in handling various NLP tasks.

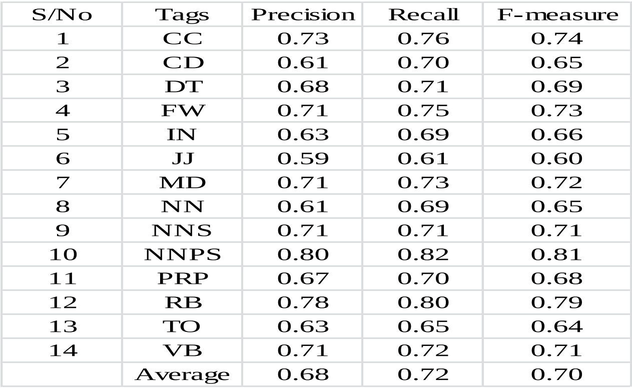
ii. NLTK is an open source tool that contains open source python source codes, Language modules, information and documentation for innovative research works in Natural Language Processing.

iii. NLTK supports numerous NLP tasks such as stemmer; tokenizer, POS tagger, classifier with compatibility for Windows, Mac, and Linux for Language such as English. Besides, it contains few distinctive corpora Yoruba.

Evaluation of the Model

Tag-wise precision recall is used to evaluate the model. Tables 2 and 3 showed the precision and recall as obtained from the test as well as the F-measure which is calculated from the relationship between recall and precision when β is 0.5.

Table 2: 50 Sentences.



S/No	Tags	Precision	Recall	F-measure
1	CC	0.73	0.76	0.74
2	CD	0.61	0.70	0.65
3	DT	0.68	0.71	0.69
4	FW	0.71	0.75	0.73
5	IN	0.63	0.69	0.66
6	JJ	0.59	0.61	0.60
7	MD	0.71	0.73	0.72
8	NN	0.61	0.69	0.65
9	NNS	0.71	0.71	0.71
10	NNPS	0.80	0.82	0.81
11	PRP	0.67	0.70	0.68
12	RB	0.78	0.80	0.79
13	TO	0.63	0.65	0.64
14	VB	0.71	0.72	0.71
	Average	0.68	0.72	0.70

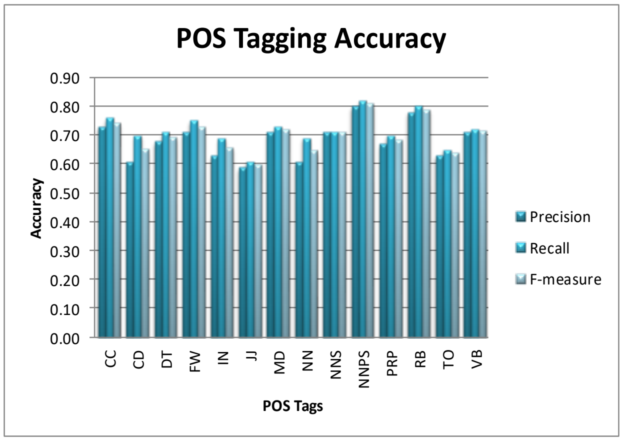


Figure 6: POS Tagging Accuracy.

Table 3: 100 Sentences.

S/No	Tags	Precision	Recall	F-measure
1	CC	0.83	0.82	0.82
2	CD	0.64	0.72	0.68
3	DT	0.71	0.79	0.75
4	FW	0.69	0.73	0.71
5	IN	0.70	0.72	0.71
6	JJ	0.63	0.69	0.66
7	MD	0.70	0.74	0.72
8	NN	0.65	0.70	0.67
9	NNS	0.72	0.76	0.74
10	NNPS	0.83	0.87	0.85
11	PRP	0.65	0.73	0.69
12	RB	0.63	0.70	0.66
13	TO	0.75	0.80	0.77
14	VB	0.76	0.83	0.79
	Average	0.71	0.76	0.73

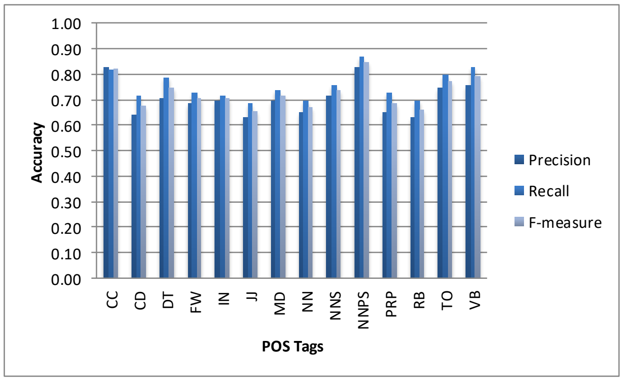


Figure 7: POS Tagging Accuracy.

Table 4: Average Precision, Recall & F-measure.

Precision	Recall	F-measure	Test Corpus Size
0.68	0.72	0.7	50 sentences
0.71	0.76	0.73	100 sentences

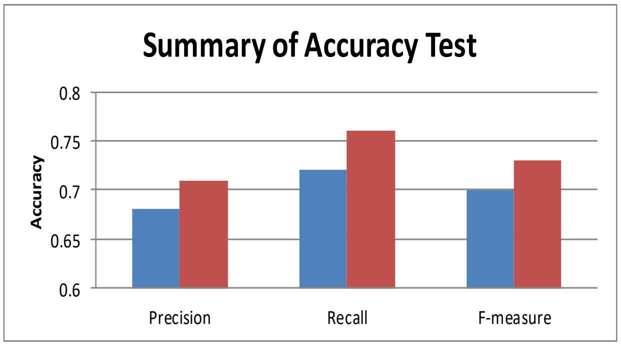


Figure 8: Summary of Accuracy Test.

DISCUSSION

One of the most generally used models in the statistical approach for POS tagging is the Hidden Markov Model. Its principle objective is to assign an optimal sequence of POS tags to a sequence of words in a given sentence. The problem of finding the optimal sequence of POS tags to a sequence of words can be done using different algorithms of which the most used one is the Viterbi algorithm. Hence, the Viterbi algorithm is adapted for this Hidden Markov Model component of the tagger in this research.

Also, NLTK and Python programming language are used as implementation tools due to their ease of application in natural language processing. They are easy to use and process text with different integrated components. There is no readily available annotated Yoruba Corpus during this research, so manual tagging was adopted. The pre-processing components (tokenizer and tagset analyzer) are developed for use by the HMM tagger. The Viterbi algorithm is implemented for finding the optimal path in the HMM algorithm based tagger.

The result of the experiment demonstrated the effectiveness of HMM as a POS tagging algorithm using very small amount of training data set.

CONCLUSION

Part of Speech tagging has gotten more significant since some companies of repute such as Google; Microsoft etc. are focusing on Natural Language Processing applications. POS tagging is playing an imperative role in various language processing and speech applications. Currently, there are numerous tools for POS tagging but little attention is drawn to Yoruba language despite the very large number of people that speak the language.

In this research, our exertion was on the modification of existing POS tagging algorithm to accommodate Yoruba Corpus. Hidden Markov Model algorithm-based tagger was adopted for modification with average precision of 0.7, recall of 0.74 and performance of 74%. HMM accuracy is dependent on the size of training data; it is therefore believed that the performance can be increased if training data size is increased.

RECOMMENDATION

There is still room for improvement as there is still a lot of work to be done in few areas. Achieving 0.7 as precision and 0.74 as recall is good for a language with a few readily available resources such as Tagset, corpus and other electronic resources to aid easy and effective development of POS taggers. The accuracy of this algorithm can be improved on by expanding the Tagset; increasing the training data size so that the algorithm can handle more ambiguous words; create cleaned Yoruba Language corpus that can serve as basis for other researchers interested in the Language. Our result and procedures can also be compared with results achieved by other POS tagging algorithms.

REFERENCES

1. Adedjouma, A.S., R.J.O. Aoga, and A.M. Igue. 2013. "Part-of-Speech Tagging of Yoruba Standard Language of Niger-Congo Family". *Research Journal of Computer Information Technology Sciences*. 1(1): 2–5.

2. Abiola, O.B., A.O. Adetunmbi, and. A. Oguntimilehin. 2013. "A Computational Model of English To Yoruba Noun-Phrases Translation System". *FUTA Journal of Research in Science*. (1): 34–43.

3. Abiola, O.B., A.O. Adetunmbi, and A. Oguntimilehin. 2015. "Using Hybrid Approach for English to Yoruba Text to Text Machine Translation System (Proposed)". Corpus ID: 212544164. 4(8): 308–313.

4. Agbeyangi, A.O. 2016. "Morphological Analysis of Standard Yorùbá Nouns". *American Journal of Engineering Reseach (AJER)*, 5(6): 8–12.

5. Agbeyangi, A.O., S.I. Eludiora, and O.A. Adenekan. 2015. "English to Yorùbá Machine Translation System using Rule-Based Approach English to Yorùbá Machine Translation System using Rule-Based Approach" *JMEST*. 2(8).

6. Alfan, Farizki, Wicaksono and Ayu Purwarianti. 2010. "HMM Based Part-of-Speech Tagger for Bahasa Indonesia HMM Based Part-of-Speech Tagger for Bahasa Indonesia". *Proceedings of 4th International MALINDO (Malay and Indonesian Language) Workshop*. 2 August 2010, (June 2014). Retrieved from https://www.researchgate.net/publication/209387036.

7. Amrullah, A.Z., R. Hartanto, and I.W. Mustika. 2017. "A Comparison of Different Part-of-Speech Tagging Techniques for Text in Bahasa Indonesia". *Proceedings - 2017 7th International Annual Engineering Seminar, InAES 2017*. https://doi.org/10.1109/INAES.2017.8068538.

8. Anish, A. 2008. "Part of Speech Tagging for Malayalam". Retrieved from http://nlp.amrita.edu:8080/project/mhrd/ms/Malayalam/ANIsh POSthesis.pdf.

9. Antony, P.J. 2011. "Parts Of Speech Tagging for Indian Languages: A Literature Survey". *International Journal of Computer Applications*. 34(8): 975–8887.

10. Barua, G. and P.K. Dutta. 2013. "An Online Semi Automated Part of Speech Tagging Technique Applied to Assamese". Indian Institute of Biology: Guwahati, India.

11. Beata, M. 2003. "Comparing Data Driven Algorithm for POS Tagging of Swedish". Centre for Speech Technology,Department of Speech, Music and Hearing, Royal Institute of Technology: Stockholm, Sweden, SE-100, 44–52.

12. Chowdhury, S.M.A., M. Nahid, U. Minhaz, I. Mohammad, M.M. Hassan, and M.E. Haque. 2004. "Parts of Speech Tagging of Bangla Sentence". *Proceeding of the 7th International Conference on Computer and Information Technology (ICCIT)*. Bangladesh.

13. Enikuomehin, O.A. 2015. "A Computerized Identification System for Verb Sorting and

Arrangement in a Natural Language : Case Study of the Nigerian Yoruba Language". *European Journal of Computer Science and Information Technology*. 3(1), 43–52.

14. Francis, M. 2014. "A Comprehensive Survey on Parts of Speech Tagging Approaches in Dravidian Languages". Shah Abdul Latif University: Sindh, Pakistan. 7–10.

15. Garrette, D. and J. Baldridge. 2012. "Type-Supervised Hidden Markov Models for Part-of-Speech Tagging with Incomplete Tag Dictionaries". *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural language Processing and Computational Natural Language Learning*. Jeju Island, Korea. 821–831.

16. Gimpel, K., N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, … and N.A. Smith. 2011. "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments". *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*. (2), 42–47. https://doi.org/10.1.1.206.3224

17. Hasan, F.M. 2006. "Comparison of Different POS Tagging Techniques for Some South Asian Languages". BRAC University: Dhaka, Bangladesh (December).

18. Hasan, F.M., N. Uzzaman, and K. Mumit. 2007. "Comparison of Different POS Tagging Techniques (N-gram, HMM and Brill's Tagger) for Bangla". *Advances and Innovations in Systems, Computing Sciences and Software Engineering*. (pp. 121–126). https://doi.org/10.1007/978-1-4020-6264-3_23.

19. Hoque, M.N., and M.H. Seddiqui. 2016. "Bangla Parts-of-Speech Tagging using Bangla Stemmer and Rule Based Analyzer". In: *18th International Conference on Computer and Information Technology, ICCIT 2015*. pp. 440–444. *https://doi.org/10.1109/ICCITechn.2015.7488111*.

20. Jurafsky, D. and J.H. Martin. 2016. "Part-of-Speech Tagging". *Speech and Language Processing*. (Chapter 10), 1–55. Retrieved from http://en.wikipedia.org/w/index.php?title=Part-of-speech_tagging&oldid=550410494.

21. Jurafsky, D. and J.H. Martin. 2018. *Speech and Language Processing*. Stanford University Press: Standford, CA.

22. Khoja, S. 2001. "APT : Arabic Part-Of-Speech Tagger". *Proceedings of the Student Workshop at NAACL*, 20--25.

23. Kumawat, D. and V. Jain. 2015. "POS Tagging Approaches: A Comparison". *International Journal of Computer Applications*. 118(6): 975–8887. Retrieved from http://research.ijcaonline.org/volume118/number6/pxc3903148.pdf.

24. Larsson, M. 1995. "Part-of-Speech Tagging Using the Brill Method". *Technology*. Corpus ID: 16704599.

25. Marco, C.S. 2012. "An Open Source Part-of-Speech Tagger for Norwegian: Building on Existing Language Resources". *IREC Conf Proceedings*. 4111–4117.

26. Mohamed, H., N. Omar, and M.J. Ab Aziz. 2011. "Statistical Malay Part-of-Speech (POS) Tagger using Hidden Markov Approach". *2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011*. (June), 231–236. https://doi.org/10.1109/STAIR.2011.5995794.

27. Onyenwe, I.E., M. Hepple, S. Sheffield, and S. Sheffield. 2014. "Part-of-Speech Tagset and Corpus Development for Igbo, An African Language" *LAW VIII: The 8th Linguistic Annotation Workshop*. Dublin, Ireland. August 23-24, 2014 93–98.

28. Padr, I.S., H. Rodr, S. Inform, U. Polit, J. Girona, C. Editor, and R. Mooney. 2000. "A Machine Learning Approach to POS Tagging". *Machine Learning*. 39: 59–91.

29. Syracuse University. 2015. "Part-Of-Speech (POS) Tagging : Feature Classification Evaluation of Results Recall HMM". Syracuse University, ischool: Syracuse, NY.

30. Zoubin, G. 2001. "An Introduction to Hidden Markov Model and Bayesian Network". *Intenational Journal of Pattern Recognition and Artificial Intelligence*. 1(15): 9–42. https://doi.org/10.1142/S0218001401000836.

SUGGESTED CITATION

Enikuomehin, A.O. and A.A. Tijani. 2020. "Implementation of Hidden Markov Model on Lagos Nigeria Women Union (NWU) Yoruba Speech Corpus". *Pacific Journal of Science and Technology*. 21(2): 214-223.

Pacific Journal of Science and Technology