https://doi.org/10.5281/zenodo.14966941 Ife Journal of Technology, Vol. 29(2), 2024

Named Entity: Extended Feature Analysis for Improved Recognition for Yorùbá

Franklin O. Asahiah, Abayomi E. Adegunlehin
Department of Computer Science and Engineering, Obafmi Awolowo University, Ile-Ife, Nigeria

ABSTRACT

The research assesses the significance of four distinct features—Part of Speech (PoS), surrounding words, prefixes, and suffixes—within a Named Entity Recognition (NER) task specifically focused on the Yorùbá language. Employing a machine learning methodology, the study utilizes a Conditional Random Fields (CRF)-based approach to construct a NER system tailored for the Yorùbá language. Additionally, the research adopts a "Leave-one-out-do-the-rest" experimental design to evaluate the influence of each feature on the recognition process. A mini PoS tagger dedicated to the Yorùbá language has been trained using CRFsuite of scikit-learn as part of this research. The dataset is systematically divided into 80% for training purposes and 20% for testing. The findings indicate that the omission of PoS tags led to a marked reduction in Recall, Precision, and F-measure. In contrast, while the exclusion of surrounding words also resulted in decreases in these metrics, the impact was notably less pronounced. Performance measured by F1 score dropped by 8.07% when the context of the words was ignored while absence of POS tags reduced performance by 7.04%. The removal of prefixes and suffixes demonstrated a relatively minor effect on the overall performance of the system. Conclusively, the research asserts that PoS tags and the words within the surrounding context emerge as the most critical features for effective NER modeling in the Yorùbá language. The integration of PoS tags appears to elucidate the enhanced performance observed, even when employing a machine learning approach that may be deemed less robust.

KEYWORDS

Named Entity
Information Extraction
Conditional Random Fields
Features

1. INTRODUCTION

Named Entity Recognition (NER) is an information extraction task that chiefly seeks to identify a Named Entity (NE) in a text and arrange or classify them into a predefined class. NER, also known as NE extraction, NE detection or NE identification became an information extraction task at the Message Understanding Conference-6 (MUC-6) in 1995 (Grishma and Sundheim, 1996). Usually, NER task is treated as a two-step process that involves identifying the named entities in a text and secondly classifying these named entities into a set of predefined classes. The classes could include person names, location, time expressions, organization, etc. NER is a core task of natural language processing and a component for many downstream applications like search engines, customer support, content recommendation, knowledge graphs and personal assistants. NER has been found to be key to many NLP applications such as Automatic Text Summarization (Nobata *et al.*, 2003), Information Extraction Systems (Toda and Kataoka, 2005), Question-Answering Systems (Mollá *et al.* 2006; Rodrigo *et al.* 2013), and so on.

Different NER approaches are given by different researchers for different domains or languages including English, Spanish, Chinese, and Japanese as surveyed in Nadeau and Sekine (2007). These methodologies can be broadly categorized into three categories namely - Rule Based approach, Machine Learning approach and Hybrid Approach. The rule-based approach depends on handwritten linguistic rule which requires immense experience and linguistic information of the specific language or domain. (Riaz, 2011, Kaur and Gupta, 2012). The disadvantage of this approach is that it provides better results for restricted domains only making it very difficult to evolve into different languages. The machine learning approach (which includes Hidden Markov Model, Maximum Entropy, Decision Tree, Support Vector Machines and Conditional Random Fields) is popularly used in NER because these approaches are easily trainable and adaptable to different domains (Benajiba *et al.*, 2009; Amarappa and Sathyanarayana, 2015, 2017). However, these techniques require large, annotated corpus for training and testing. Hybrid approaches take the advantage of both the rule-based and machine learning based techniques. (Srikantha and Murthy, 2008; Kumar and Kiran, 2008).

Yorùbá is regarded as one of the major well-spoken languages in Nigeria and a few other African countries (Barber 2015). Eberhard *et al.*, (2019) revealed that Yorùbá is the third most spoken indigenous language in Africa after Swahili and Hausa with over 35 million native speakers. Yorùbá has several dialects but according to Asahiah *et al.*, 2017, the written language was standardized by the 1974 Joint Consultative Committee on Education. However, compared to other languages like English, European languages, Chinese, Japanese, Korean and other foreign languages like Arabic, Yorùbá language has not received much attention being a low-resource language. This is partly as a result of paucity of well-labelled training data for most NLP applications.

Numerous works have been carried out on NER but most have focused only on getting improvement by utilizing newer and more efficient machine learning algorithms. The impact of choice and availability of features used has not been serious investigated except in few instances. Furthermore, no existing work has investigated the issue of feature engineering for Yorùbá NER that could be crucial to performance improvement. The aim of this work is to investigate the possible set of features that can be used in performing NER for Yorùbá language and reveal the impact of each of these features on the recognition process. While other previous works carried out automatic feature selection (Le and Tang, 2013), This work seeks to measure the impact of each selected feature in the final performance of the model. We started by identifying the optimal set of features for the Yorùbá Named Entity Recognition task. We explored the feature sets that can be used in performing Named Entity Recognition on Yorùbá text using a machine learning approach. A comparison analysis was performed to show the impact of these features on the two-step process - identification and classification of Named Entities.

Corresponding Author: F. O. Asahiah (*sobusola@oauife.edu.ng*)
Received 9 August 2024 | **Received in revised form** 28 October 2024 | **Accepted** 21 November 2024 | **Available online** 10 February 2025
The editor responsible for coordinating the review of this article and approving its publication was *I. P. Gambo*
P-ISSN: 1115-9782 **e-ISSN:** 2536-6807

2. RELATED WORKS

Adeyemi (2016) in his study examined the relationship between Arabic and Yorùbá languages. He was able to analyse the similarity in pronunciation, comparison in construction of words between them and similarity in the use of some words in both languages with the views of some Arabic scholars.

Ikechuckwu *et al.*, (2019) in their work titled "A First Step Towards the Development of Yorùbá Named Entity Recognition System", the authors focused on creating a ground for researchers to develop a robust NER for Yorùbá language. It involved the use of some widely used features such as PoS of a word and its surrounding words, affixation, capitalization, etc., and sequence modeling framework of other language for Yorùbá in order to investigate their usefulness.

This work builds on the conditional probabilistic approach to NER using Conditional Random Field (CRF) classifier. Ekbal and Bandyopadhyay, (2009) introduced a CRF-based approach for Named Entity Recognition in Bengali and Hindi. Prefix, suffix, tags of previous words, PoS tags, first word, length of the word and gazetteer lists are some language inherited features described in this paper.

Mo *et al.*, (2017) performed NER using conditional random fields (CRF) on Myanmar language. The work was with the intention of inducing name entities automatically in Myanmar scripts and to develop a base line NER.

Vijayakrishna and Sobha (2008) proposed a Tamil Named Entity Recognition system that is domain-focused. The system handles nested tagging of named entities in the tourism domain. They were able to experiment with a CRF model by training the noun phrases of the training data.

Tkachenko and Simanovsky (2012) in their paper titled "Named Entity Recognition: Exploring Features" presented research on the complete features used in identifying supervised-based NER task, various combinations of these features and evaluation of the performance. The work tries to reveal the effectiveness of clustering features and their combinations on NER. Benajiba, *et al.*, (2007) and Benajiba and Rosso (2007) developed an Arabic Named Entity Recognition system where they used language-independent features such as contextual words, prefix and suffix information, and digit features. Benajiba and Rosso (2008) went further by combining Language-independent and Arabic-specific features in the CRF model, including POS tags, gazetteers, and nationality. The CRF based system achieved best results when all the features were combined.

Benajiba, *et al.*, (2008b) in their work examined the impact of various types of features on the different types of NE. They examined the lexical, contextual, morphological, gazetteer, and shallow syntactic features, to form 16 specific features in total. An approach involving the combination of SVM and CRF models was used with a voting scheme to rank the features. A CRF-based Arabic NER system was developed by Abdul-Hamid and Darwish (2010) using a set of simplified features for recognizing three NE types: person, location, and organization. These features include character n-grams, word n-gram, word sequence features, and word length. Adegunlehin *et al.* (2019) investigated the contribution of the Part of Speech tags and surrounding words in correctly determining if a word is a named entity and what category of named entity it belongs to for Yorùbá text. The authors showed that performance measured by F1 score dropped by 8.07% when the context of the words was ignored while absence of POS tags reduced performance by 7.04%. The performance was compared against the experimental setup in which full feature set were used. Situmeang (2022) on the other hand, indicated that careful preprocessing of text data contributes to Indonesian NER performance.

3. METHODOLOGY

This research aims to investigate and reveal the relevance of each possible sets of features in recognizing (identifying and classifying) a NE type for a Yorùbá text. The investigative approach used was to review related works on other languages that share similarities with the Yorùbá language. This is with the belief that the features sets used in developing a NER system for such languages can be adopted for Yorùbá language. According to Adeyemi (2016), parts of Yorùbá language emanated from Arabic, which has root in Semitic language, therefore establishing the lexical similarities in Yorùbá and Arabic language. With this knowledge, we reviewed works on NER for Arabic language to have a conceptual view of possible sets of features that could be applicable in developing a NER system for Yorùbá language. For this research, the following set of features have been selected; Context words, Parts of Speech, Prefixes, Suffixes.

To reveal the relevance of each of these features, we developed a NER system for Yorùbá language using a Conditional Random Fields (CRF) based machine learning technique. CRF are undirected graphical models (Lafferty *et al.*, 2001) used to calculate the conditional probability of values on designated output nodes, given values on other designated input nodes (Wallach, 2004).

Our experimental setup is dubbed "Leave-one-out-do-the-rest". This setup involves using all the features considered relevant, based on previous study and measure the performance of the classifier. We then subsequently leave one of the identified features out of the feature set and running the experiment and then measure the performance of the classifier again. This was done for every feature in the feature set, keeping every other feature in. The performance of the classifier without the inclusion of a particular feature is approximated to reflect the impact of that feature in the experiment.

The classifier (CRF) in sklearn CRFSuite used the lbfgs (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) optimization algorithm with a maximum iteration of 100 and hyperparameters c1 (coefficient for L1 regularization) and c2 (coefficient for L2 regularization) both set to 0.5.

3.1. Data Collection and Annotation

The corpus used for this research was manually annotated for PoS tags and Named Entity tags. The Named Entity categories considered are Person, Location and Time Entities. A total of 420 sentences with a total of 9,421 words were collected. A sample is shown in Table 1.0 is for a sentence that has several Person (B-Pers, I-Pers) tags.

Table 1: A Sample of the format of the training data

Sentence Number	Word	PoS		Tag
Sentence 383	Ègbọ́n	NN	# Noun	O
	Ilẹ́sanmí	NNP	# Proper Noun	B-Per
	Ní	VB	# Verb	O
	Àdùnní	NNP	# Proper Noun	B-Per
	Ìyá	NN	# Noun	O
	Dúró	NNP	# Proper Noun	B-Per
	Orímóògùnjẹ́	NNP	# Proper Noun	I-Per
	.	.	# Sentence-final Punctuation	O

In this dataset, there are 274 Persons named entity, 201 Locations named entity, and 214 Time named entity spread over 938 words. The choice to focus on the named entities PERSON, LOCATION and TIME was informed by the size of the gathered data and the paucity of the other entities in sufficient quantity for effective training. Mentions of other named entities such as Organization, Concept, Facility, etc, contained in the corpus were labeled as Others (O) since they are not covered in the scope. This is done to increase the performance of the system such that the set of features identified will be adequate in identifying and classifying the desired Named Entities (NE). Some named entity expressions are multi-word named entities (or multi-word expressions), to capture this we used Inside-

Outside-Beginning (IOB) notation for the class labels so as to correctly identify the named entity span.

Crucial to the annotation process is resolving annotation discrepancies and reaching reasonable agreement to achieve uniform annotation across the corpus. During annotation, ambiguity of words was handled by giving strict consideration to the context of a word when deciding its NE category. To address this issue named entities were annotated considering "Function over Form" annotation manner, thus the same named entity could be tagged differently in a different context, like the following example; "*... ní àdúgbò Fájuyì*"; it is a fact that "*Fájuyì*" is actually a person's name but here it's preceded by a location designator (*àdúgbò*). So, this string of words won't be tagged in isolation but will be tagged according to the designator identity which is a location.

The following assumptions were made in all using MUC-7 guidelines:

i. Time expressions that are not specific about the actual time are not considered. For instance, mentions like *alé, ìrọ̀lẹ́, òsán, àárọ̀, òwúrọ̀, ìdájí*, etc, are not considered as a Time named entity because "*alé*" refers to any time between 6 pm and 11:59 pm. It would be more specific if "*alé*" can be used in the form "*agogo mewaa alé*" (10:00 pm).

ii. The word "*Yorùbá*" could be a Concept entity, Location entity, Geo-political entity or Language entity; but for the purpose of this research, mentions of this word in the corpus were annotated as Others (O) in other to strip off ambiguity.

Data used for this research was prepared in CSV (Comma Separated Values) format using Microsoft Excel package. The created data file contained one token per line with a full stop '.' representing each sentence boundary. The corpus was annotated at the sentence level and then was prepared into four columns as shown in Table 2, wherein the first column is the 'Sentence number', the second column is the 'Word', the third column is the 'Part of Speech' tag of that word, and the fourth column is the 'Named Entity' tag of that word.

Table2: Processed data sample

Sentence Number	Word	PoS		Tag
Sentence 200	Lehin	IN	# Preposition	O
	ọdún	NNP	# Proper Noun	B-Tim
	mẹwa	NNP	# Proper Noun	I-Tim
	ìdájọ́	NN	# Noun	O
	bẹ̀rẹ̀	VB	# Verb	O
	Fún	IN	# Preposition	O
	Dúró	NNP	# Proper Noun	B-Per
	Orímóògùnjé	NNP	# Proper Noun	I-Per

3.2. Part-of-Speech tagger for *Yorùbá* Language

Since there is no standalone PoS tagger application for *Yorùbá* language which suits for this research, therefore, a mini dedicated *Yorùbá* PoS tagger was trained using Conditional Random Field. Sample of the data used is shown Table 3.

3.3. Recognition Technique

Conditional Random Field (CRF) has shown success in various sequence modeling tasks including NER tasks (Sha and Pereira 2003). Among CRF toolkits, CRF++ and CRFsuite are the most popular choices. However, CRFsuite which is more robust and faster to train was chosen for this work. As with many NLP tasks, we will be working with sequence data. We consider sentences to be a sequence in which each word's meaning is dependent on both the other words in the sentence and the order in which they appear. CRFs aims to find y that maximize p(y|x) for the sequence x; where y is the labeled sequence, and p(y|x) is the probability of y given x, considering the previous and the succeeding elements.

Table 3: Data for POS tagger development

Word	PoS	Description (not included in data)
Ẹ̀gbọ́n	NN	# Noun
Ilẹ́sanmí	NNP	# Proper Noun
Ní	VB	# Verb
Àdùnní	NNP	# Proper Noun
Ìyá	NN	# Noun
Dúró	NNP	# Proper Noun
Orímóògùnjé	NNP	# Proper Noun
.	.	# Sentence-final Punctuation

Unlike Maximum Entropy, CRF does not require careful feature selection in order to avoid overfitting. CRF being conditionally trained has the freedom to include arbitrary features, non-independent features, and the ability to automatically construct the most useful feature combinations by feature induction. It requires training and a testing data set. From the annotated corpus 80% of the data set is used as train set and 20% used as test set.

We used linear chain CRFs where **p(y|x)** is defined in equation 1 by Lafferty et. al. (2001):

$$P_{(y|x)} = \frac{1}{z_{(x)}} \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_j f_j(s_{i-1}, s_i, x, i)\right\} \quad [1]$$

where $\boldsymbol{f_j(s_{i-1}, s_i, x, i)}$ is the function for the properties of transition from the state $\boldsymbol{s_{i-1}}$ to $\boldsymbol{s_i}$ with the input $\boldsymbol{x}$; $\boldsymbol{\lambda_j}$ is the parameter optimized by the training; $\boldsymbol{s_{i-1}}$ is the previous state; $\boldsymbol{s_i}$ is the current state; **y** is sequence of labels; $\boldsymbol{x}$ is observed input word sequence (Sentence); $\boldsymbol{m}$ is the number of feature templates; **n** is the length of the sentence.

When applying CRFs to the NER problem, an observation sequence is a token of a sentence and the state sequence is its corresponding label sequence. A feature function $\boldsymbol{f_i(s_{i-1}, s_i, x, i)}$ has a value of 0 for most cases and is only set to be 1, when $\boldsymbol{s_{i-1}, s_i}$ are certain states and the observation has certain properties. For Yorùbá NER, one possible feature function could measure how much we suspect that the current word should be labeled as LOCATION given that the previous word is "***ni***". Therefore, $\boldsymbol{f_1(x, i, s_i, s_{i-1})} = 1$ if s_i = LOC, the preceding word is '***ni***' and the word is in sentence case; 0 otherwise. If the weight λ_1 associated with this feature is large and positive, then this feature is essentially saying that we prefer labelings where words in sentence case that have '***ni***' as its preceding word get labeled as **LOCATION**.

3.4. Features Experiment

The purpose of this experiment is to evaluate the relevance of each of the four features identified in the recognition of each of the Named Entity categories identified in the data set. Therefore, to carry out the experiment, all the four features are used in developing the system and evaluated. The result of the evaluation was used as a baseline for benchmarking the main experiment for this research. The standard evaluation metrics used are Precision, Recall and F-measure. The baseline experiment in which all the feature combinations were used gave a Recall value of 88.16, a Precision value of 89.92 and a F-measure value of 89.06%. This is presented and discussed later in Table 10.

4. RESULT AND DISCUSSION

4.1. Relevance of Parts of Speech tags

In this case, PoS of the tokens as a feature were removed but other features were considered. In Table 4, the confusion matrix showed the classifier misclassified more of PERSON entity, followed by LOCATION entity as a non-entity (OTHERS), while only a few words identified as an entity were wrongly classified into another entity class. The reason for this could be because the classifier attached more weights on PoS tags to recognize the PERON entity and LOCATION entity. Other features such as surrounding words, capitalization is not enough to identify a word as a Person entity or Location entity. Therefore, it could be assumed that the use of PoS tags is a very necessary feature for identifying a word that belongs to PERSON and LOCATION entities in a *Yorùbá* text.

The parameters calculated from Table 4 and shown in Table 5 indicated that compared to the baseline, Recall fell by a big margin of 12.3% while Precision reduced marginally by 0.67% and F-measure reduced by 7.04%. The implication of this result for the absence of the POS from the feature set has a very strong impact on the Recall and hence, on the F-measure, but significantly less on the Precision in NER of *Yorùbá* sentences.

Table 4: Confusion Matrix for performance of the Yorùbá NER without Part-of-Speech tags

		PREDICTED LABEL						
		B-Loc	B-Per	B-Tim	I-Loc	I-Per	I-Tim	O
TRUE LABEL	B-Loc	**138**	8	5	1	1	0	48
	B-Per	13	**165**	0	0	9	0	87
	B-Tim	0	0	**212**	0	0	0	2
	I-Loc	5	0	0	**15**	1	5	12
	I-Per	0	3	0	0	**115**	0	17
	I-Tim	0	0	0	0	0	**209**	4
	O	5	16	3	2	2	6	8313

Table 5: Performance of the Yorùbá NER system without Part-of-Speech tags

Metrics	Simple Average	Weighted Average
Recall	0.7704	0.7588
Precision	0.8955	0.8925
F-measure	0.8283	0.8202

4.2. Relevance of Context Words

In this case, all features except the surrounding words were used. The surrounding words are the word immediately preceding the present token and the word immediately after the present token.

The confusion matrix shown in Table 6, indicated that the classifier misclassified many words that are non-entities (OTHERS) as TIME entities, while some words that are LOCATION entity were wrongly classified as PERSON entity. This is because of words that can be used as Person's name and as well as a Location name. Therefore, it could be assumed that the use of surrounding words as a feature in *Yorùbá* Named Entity Recognition task helps the classifier to correctly distinguish between such cases.

The NER performance of the classifier shown Table 7 measured by Recall, Precision and F-measure dropped by 7.19%, 9.34% and 8.27% respectively from the baseline. This result reveals that for *Yorùbá* Named Entity recognition task, words surrounding a particular word gives a lot of information in predicting the Named Entity tag of that particular word.

Table 6 Confusion Matrix for the performance of the Yorùbá NER without surrounding words

		PREDICTED LABEL						
		B-Loc	B-Per	B-Tim	I-Loc	I-Per	I-Tim	O
TRUE LABEL	B-Loc	**130**	29	13	1	3	1	24
	B-Per	21	**233**	2	0	7	2	9
	B-Tim	2	2	**205**	0	0	2	3
	I-Loc	8	0	0	**7**	8	12	3
	I-Per	0	2	0	0	**131**	2	0
	I-Tim	1	0	2	1	2	**202**	5
	O	0	4	33	7	13	33	**8257**

Table 7: Performance of the Yorùbá NER system without surrounding words

Metrics	Simple Average	Weighted Average
Recall	0.8054	0.8099
Precision	0.8027	0.8058
F-measure	0.8041	0.8079

4.3. Performance of the *Yorùbá* NER system without Prefixes

In the third case experiment, all the features except the Prefixes were used and Recall, Precision and F-measure presented in Table 8 showed a relative decrease of 0.93%, 1.17% and 1.04% respectively compared to the baseline. The impact of prefixes as a feature is relatively weaker than that of POS and Context words. It, nevertheless, was useful in disambiguating some words

Table 8: Performance of the Yorùbá NER system without prefixes

Metrics	Simple Average	Weighted Average
Recall	0.8730	0.8723
Precision	0.8878	0.8875
F-measure	0.8803	0.8798

4.4. Performance of the *Yorùbá* NER system without Suffixes

In this experiment, all the features were used except the suffixes of the tokens. The result as shown in Table 9 showed a relative decrease of 0.84%, 0.55% and 0.73% respectively compared to the baseline. This shows that the use of the suffix of a particular word did not significantly affect the performance of the system unlike the absence of prefix in the feature sets.

Table 9: Performance of the Yorùbá NER system without Suffixes

	Simple Average	Weighted Average
Recall	0.8736	0.8732
Precision	0.8944	0.8937
F-measure	0.8839	0.8833

5. OVERVIEW OF THE EXPERIMENTAL RESULT

A general overview of the experimental result on the relevance of each feature on the performance of the model is shown in Table 9 and Figure 1. Figure 1 showed that the most impactful of the features are, in order of importance, the POS tag, the words in surrounding context, prefixes and suffixes. The research has revealed that each feature used either helps to identify a word as a possible named entity and/or helps to classify such word to the proper named entity class it belongs.

The exclusion of POS resulted in the system to have the highest False Negatives which means the system could not

identify some words as a possible named entity, as a result some named entity words are classified as non-entity.

Table 10: Details of all the experimental results

Metrics	All feature combination	Without PoS	Without Context Words	Without Prefixes	Without Suffixes
TP	968	854	908	963	963
FP	92	85	213	103	98
FN	105	221	167	112	112
Recall	88.16%	75.88%	80.99%	87.23%	87.32%
Precision	89.92%	89.25%	80.58%	88.75%	89.37%
F-measure	89.06%	82.02%	80.79%	87.98%	88.33%

On the other hand, exclusion of surrounding-words (Context Words) from the feature set resulted in the system to have a high False Positive which means that some words identified as a possible named entity could not be correctly classified into their respective named entity class labels. The exclusion of surrounding-words from all the feature combination resulted in the system having the highest number of misclassifications (213-False Positives and 167-False Negatives) followed by exclusion of PoS from the feature set which gave 85 False Positives and 221 False Negatives. The exclusion of Prefixes and Suffixes from the feature sets did not significantly reduce the system's performance by much.

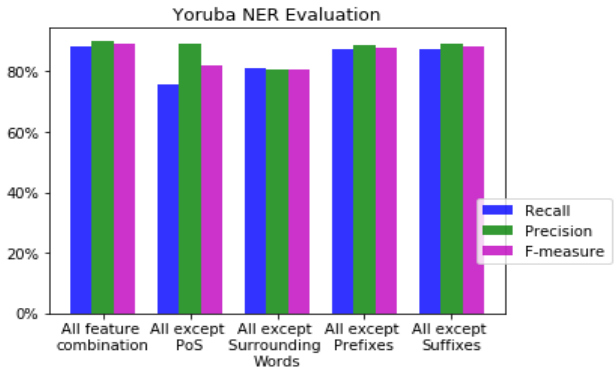

Figure 1: Yorùbá NER Model performance vs Features removal

6. CONCLUSION

This research has been able to show that the two most important features in NER modeling for Yorùbá language are POS tags and words in surrounding context. Both of them significantly affect whether a word is recognized as a named entity or not and if recognized, how correct the classification will be. Compared to the reported performance for Yorùbá in Oyewusi *et al.* (2021), the inclusion of POS tags might be able to explain the better performance despite using a machine learning approach that is considered less powerful It is therefore paramount to devout more research to developing a more accurate POS tagger with finer granularity. In addition, word embedding should be investigated as an alternative in n-gram model of word context.

REFERENCES

Abdul-Hamid, A. and Darwish K., (2010). Simplified Feature Set for Arabic Named Entity Recognition. *In Proceedings* of the 2010 Named Entities Workshop (NEWS 2010), pages 110–115, Stroudsburg, PA.

Adegunlehin A. E., Asahiah F. O. and Onifade M. T. (2019) Investigation of Feature Characteristics for Yorùbá Named Entity Recognition System. In Proceedings of Application of Information and Communication Technologies to Teaching, Research, and Administration 2019: 108-111, (Nigeria)

Adeyemi, K. (2016) A Study of Relationship Between Arabic and Yorùbá Languages. Open Journal of Modern Linguistics, 6, 219-224. doi: 10.4236/ojml.2016.63023.

Amarappa S. and Sathyanarayana S. V., (2015). "Kannada Named Entity Recognition and Classification using conditional Random Fields," 2015 International Conference on Emerging Research in Electronics, Computer Science and Technology (ICERECT), Mandya, India, 42 pp. 186-191, doi: 10.1109/ERECT.2015.7499010.

Amarappa S., and Sathyanarayana, S. V. (2017). Kannada Named Entity Recognition and Classification using Support Vector Machine. Transactions on Machine Learning and Artificial Intelligence, 5(1), 43. https://doi.org/10.14738/tmlai.51.2549

Asahiah, F. O., Odejobi, O. A., & Adagunodo, E. R. (2017). Restoring Tone-Marks in Standard Yorùbá Electronic Text: Improved Model. Computer Science, 18(3). pp. 301–315 https://doi.org/10.7494/csci.2017.18.3.2128

Barber, K. (2015). Yorùbá Language and Literature". In Oxford Bibliographies Online in African Studies. Retrieved July 12, 2019 from https://www.oxfordbibliographies.com/view/document/obo-9780199846733/obo-9780199846733-0156.xml

Benajiba, Yassine and Paolo Rosso. (2007). ANERsys 2.0: Conquering the NER task for the Arabic language by combining the maximum entropy with POS-tag information. In Proceedings of Workshop on Natural Language-Independent Engineering, 3rd Indian International Conference on Artificial Intelligence (IICAI-2007), pages 1,814–1,823, Mumbay.

Benajiba, Yassine, Paolo Rosso, and Jos´e Miguel Bened´ı Ruiz. (2007). ANERsys: An Arabic named entity recognition system based on maximum entropy. In Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2007), pages 143–153, Berlin.

Benajiba, Yassine and Rosso P. (2008). Arabic named entity recognition using conditional random fields. In Proceedings of the Workshop on HLT & NLP within the Sixth International Conference on Language Resources and Evaluation (LREC 2008), pages 143–153, Marrakech.

Benajiba, Yassine, Diab M., and Rosso P. (2008b). Arabic named entity recognition using optimized feature sets. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008), pages 284–293, Stroudsburg, PA.

Benajiba, Y., Diab, M., and Rosso, P. (2009). Arabic Named Entity Recognition: A Feature-Driven Study, IEEE Transactions on Audio, Speech, and Language Processing, 17(5), pp 926-934.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). (2021). Ethnologue: Languages of the World. Twenty-fourth edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Ekbal, A., Bandyopadhyay, S. (2009). A Conditional Random Field approach for named entity recognition in Bengali and Hindi, Germany: Department of Computational Linguistics, University of Heidelberg, India: Department of Computer Science and Engineering Jadavpur University.

Grishman, Beth Sundheim. (1996). Message Understanding Conference-6: "A Brief History". In the proceedings of the 16th International Conference on Computational Linguistics (COLING), pages 466-471, Center for Sprogteknologi, Copenhagen, Denmark

Ikechukwu I, Adebayo O, and Bosede A. (2019). A First Step Towards the Development of Yorùbá Named Entity Recognition System. International Journal of Computer Applications. 182. 1-4.

Kashif Riaz. (2010). "Rule-based Named Entity Recognition in Urdu". Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 126 – 135, Uppsala, Sweden.

Kaur Kamaldeep; Vishal Gupta. 2012 "Name Entity Recognition for Punjabi Language". International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN: 2249-9555 Vol. 2, No.3.

Kumar, P. P. and Kiran, V. R. (2008), A Hybrid Named Entity Recognition System for South Asian Languages, In Proceedings of

the IJCNLP-08 workshop on NER for South and Sound East Asian Languages, pp. 83-88.

Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: ICML 2001 Proceedings of the Eighteenth International Conference on Machine Learning, pp. 282–289

Le, Huong and Tran, Luan. (2013). Automatic feature selection for named entity recognition using genetic algorithm. ACM International Conference Proceeding Series. 81-87. 10.1145/2542050.2542056.

Mo H. M., Nwet K. T., Soe K. M. (2017). CRF-Based Named Entity Recognition for Myanmar Language. In: Pan JS., Lin JW., Wang CH., Jiang X. (eds) Genetic and Evolutionary Computing. ICGEC 2016. Advances in Intelligent Systems and Computing, vol 536. pp 204-211 Springer, Cham. https://doi.org/10.1007/978-3-319-48490-7_24.

Mollá, D., Van Zaanen, M. and Smith, D. (2006). Named Entity Recognition for Question Answering: In Proceedings of the Australasian Language Technology Workshop (ALTW2006), pp. 51–58.

Nadeau, D. and Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. Lingvisticae Investigationes, 30(1): 3-26.

Nobata, C. Sekine, S. and Isahara, H. (2003). Evaluation of Features for Sentence Extraction on Different Types of Corpora. Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering. 12 pp 292-36.

Oyewusi, W.F., Adekanmbi, O., Okoh, I., Onuigwe, V., Salami, M.I., Osakuade, O., Ibejih, S., and Musa, U.A. (2021). NaijaNER: Comprehensive Named Entity Recognition for 5 Nigerian Languages. ArXiv, abs/2105.00810.

Rodrigo, Á, Pérez-Iglesias, J., Peñas, A., Garrido, G. and Araujo, L. (2013). Answering Questions About European Legislation. Expert Systems with Applications, 40(15): 5811-5816

Sha, F. and Pereira, F. (2003). Shallow Parsing with Conditional Random Fields. In Conference on Human Language Technology and North American Association for Computational Linguistics (HLT-NAACL), pp. 213–220.

Srikantha P. and Murthy K. N. (2008), Named Entity Recognition for Telugu. In Proceedings of IJCNLP-08 workshop on NER for South and Sound East Asian Languages. pp. 41-50.

Tkachenko, M., and Simanovsky, A. (2012). Named Entity Recognition: Exploring Features. In Proceedings of the 11th Conference on Natural Language Processing (KONVENS 2012), Vienna, Austria, pp 118-127.

Toda, H. and Kataoka, R. (2005). A search result clustering method using informatively named entities. Proceedings of the Seventh Annual ACM International Workshop on Web Information and Data Management, ACM. pp. 81-86.

Vijayakrishna, R., Sobha, L., (2008). Domain focused named entity recognizer for Tamil using conditional random fields. In: Proceeding of the IJCNLP-08 Workshop on NER for South East Asian Languages.

Wallach, H. M. (2004). Conditional random fields: an introduction. University of Pennsylvania CIS Technical Report MS-CIS-04-21, 24.