Thomas Adewumi University

Journal of Innovation,

Science and Technology (TAU-JIST)

ISSN: 3043-503X

RESEARCH ARTICLE

A NAMED ENTITY RECOGNITION SYSTEM FOR BASSA, EBIRA, AND OKUN LANGUAGES

Catherine Dupe Omidiji, Emeka Ogbuju, Joshua Jimba, Francisca Onaolapo Oladipo

Department Of Computer Science, Federal University Lokoja. Nigeria.

Corresponding Author Email: catherine.omidiji-msc@fulokoja.edu.ng

ARTICLE DETAILS

Article History:

Received 02 July 2024
Accepted 05 October 2024
Available online 10 November 2024

ABSTRACT

This study focuses on developing a Named Entity Recognition (NER) system tailored specifically for low-resource languages spoken in Nigeria, namely Bassa, Ebira, and Okun. It sheds light on the crucial role of NER in natural language processing, underscoring the challenges encountered in creating NER systems for languages with limited resources, such as annotated data and linguistic tools. Its objective is to bridge this gap by offering a comprehensive overview of NER systems designed for these three Nigerian languages. The discussion delves into various approaches, hurdles, and recent progressions within the field. It stresses the significance of accurately identifying words for diverse language-processing tasks and emphasizes the meticulous process of collecting data, particularly text documents, in Bassa, Ebira, and Okun. Additionally the study uses machine learning approaches such as deep neural network (spaCy) based on Convolutional Neural Network (CNN) and Conditional Random Fields (CRF) which play an important role in the proper identification of named entities. Highlights the potential of recent advancements in machine learning and natural language processing to improve NER systems for languages facing resource constraints. This advancement not only enhances the accuracy and precision of NER but also advocates for the inclusivity and accessibility of NLP technologies on a global scale. The outcome of this endeavor manifests itself in promising results, with an impressive accuracy of 0.98, an F1-Score of 0.98, and a precision of 0.97 across all three languages.

KEYWORDS

Globalisation; NLP, NER, Recognition, Bassa, Okun, Ebira, SVM

Quick Response Code	Access this article online	
	Website: https://journals.tau.edu.ng/index.php/tau-jist	**DOI:** https://doi.org/10.5281/zenodo.14201543

Introduction

Natural Language Processing (NLP) is a field of computer science that focuses on the interaction between human language and computers, to enable machines to understand, interpret, and generate human language (Akindele *et al*., 2022). NER plays a vital role in various NLP applications, such as text summarization, machine translation, and information retrieval. However, most NER systems are designed for major languages such as English, French, and Chinese, and little research has been done on developing NER systems for low-resource languages such as Nigerian languages. Nigeria is home to over 512 ethnic languages, making it the country with the fourth-highest number of spoken languages in the world (Eberhard, 2021).

Despite obvious linguistic diversity, the majority of NLP research in Nigeria has focused on the English language, which is the country's official language (Akindele *et al*., 2022) This study focuses on three languages in Nigeria; they are Bassa, Ebira, and Okun (BEO languages). Bassa is a Niger-Congo language spoken primarily in Nigeria, specifically in the Middle Belt region, particularly in Plateau and Nasarawa states. It belongs to the Benue-Congo language family and is part of the larger Plateau subgroup (Ilọri *et al*., 2021). The Bassa language has unique linguistic features and is spoken by a significant number of people in the region. Ebira is also a Niger-Congo language spoken in Nigeria. It is primarily spoken by the Ebira people, who are found in the central part of Kogi State, as well as in some neighboring states such as Edo, Ondo, and Niger. Ebira belongs to the North-Central branch of the Benue-Congo language family and has distinct linguistic characteristics (Bendor-Samuel, 2003). Okun is another Niger-Congo language spoken in Nigeria. It is primarily spoken in the southwestern part of Kogi State, which is located in the central region of Nigeria. Okun belongs to the Yoruba branch of the Benue-Congo language family and shares some similarities with the Yoruba language, which is spoken by a larger population in the region (Aremu *et al*., 2011). Developing a NER system for these Nigerian languages contributes to the accuracy and effectiveness of NLP applications in Nigeria.

The development of an NER system for these Nigerian languages improves the accuracy and efficacy of NLP applications in Nigeria. However, several challenges complicate this development. One important challenge is the lack of annotated data, which is essential for training NER models. Unlike major languages, where extensive labelled datasets are available, low-resource languages often lack the necessary linguistic resources, making it difficult to create accurate and reliable NER systems. Additionally, the linguistic complexities of these languages, including dialectal variations, polysemy, and rich morphology, present further challenges. These complexities necessitate the development of tailored approaches and models that can handle the specific nuances of each language.

The objective of NER is to assign labels to entities like names of individuals, locations, organizations, and other relevant entities within textual documents. NER can be approached through three main methods: lexicon-based, rule-based, and machine learning-based. However, a NER system can incorporate elements from multiple categories (Keretna *et al*., 2014). Some NER approaches utilize Part-of-Speech (POS) tagging. Additionally, NER serves as a preprocessing step for tasks like information extraction or relationship extraction (Jiang *et al*., 2016). Table 1 presents a list of tools employed for NER tagging. These tools primarily rely on statistical techniques. Stanford's CoreNLP NER (also known as CRFClassifier) utilizes linear chain Conditional Random Fields (CRFs). Apache OpenNLP employs Maximum Entropy (ME). CogComp-NLP is a collection of tools, and its NER component utilizes hidden Markov models, multilayered neural networks, and other statistical methods.

Table 1. Tools for Named Entity Recognition

Library Name	PL	License
SpaCy	Python	MIT
GATE	Java	LGPL
OpenNLP	Java	Apache 2.0
CoreNLP	Java	GPL 3.0
NLTK	Python	Apache 2.0
CogcompNLP	Java	Research

The NER pipeline typically involves several stages. First, the data undergoes preprocessing, including tokenization and sentence splitting. Next, feature extraction takes place, followed by the application of machine learning models to assign tags to the data. Finally, post-processing is applied to address any tagging inconsistencies. Figure 1 provides an illustration of this pipeline.

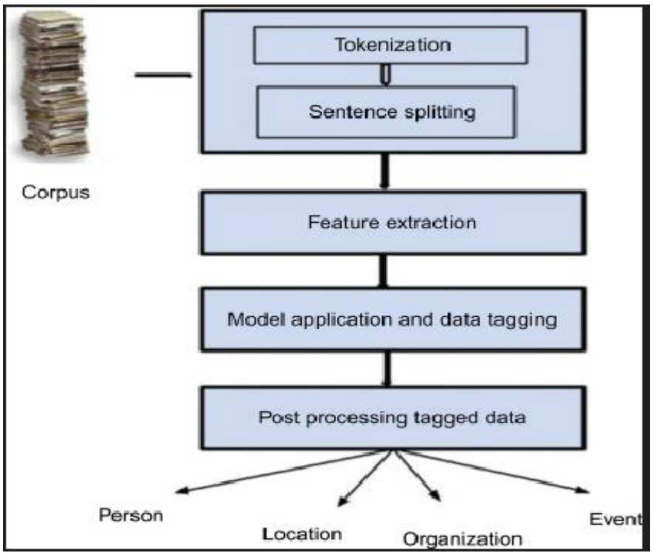


Figure 1. Pipeline of Name Entity Recognition (Kannan *et al.*, 2016)

Related Literature

There is a vibrant research thrust in Asian languages addressing NER problems. Shaalan and Raza (2019) developed a rule-based Named Entity Recognition System (NERA) for Arabic. Their approach includes incorporating local grammar, a name dictionary, and a filtering technique. The filtering technique is used to improve the system's accuracy by discarding incorrectly named entities. The evaluation of NERA resulted in F-scores of 87.7% for person, 85.9% for location, 83.15% for organization, and 91.6% for date entities. Likewise, Zaghouani (2012), developed an Arabic Named Entity Recognition System (ANERS) using rule-based techniques. The system utilizes a combination of handwritten local patterns, linguistic dictionaries, and language-specific rules to identify four types of named entities: person, organization, location, and miscellaneous. The dataset used for evaluation is called ArabiCorpus and comprises 68,943,447 Arabic words from diverse sources like newspapers, the Quran, and Arabic literature. The system's performance is compared to Benajiba's ANerCorp and Lingpipe, and it achieved the highest Precision value of 71.18% for the person entity and the best F-score value of 87.63% for the location entity. One of the applications of NER is automatic text summarization. In their study, Gupta and Lehal (2011) developed a NER System for the Punjabi language, which was subsequently integrated into an Automatic Text Summarization System. Since Punjabi is a language with limited resources, the authors manually created various gazetteer lists, such as proper name, prefix, middle name, last name, and suffix lists, based on the Punjabi corpus. These lists served as linguistic resources for the conditional rule-based technique used to extract named entities. The system achieved a Precision of 89.32%, Recall of 83.4%, and F-score of 86.25% in terms of results. However, the system reported an error rate of 13.75%, which was attributed to the system not considering proper nouns as general nouns and the unavailability of certain words in the gazetteers. In the same vein, Khan *et al., (2022)* proposed a NER approach for the Urdu language using conditional random fields. They proposed novel features and feature functions and also created the UNER-I dataset for evaluation purposes. The results indicated improved performance compared to the baseline techniques.

Manuel, (2022) Analyzed the performance of various NER models in the context of African languages. The research on community in Africa is also providing resources for advancing development in the languages across the continent. The study focuses on the quality and quantity of datasets and evaluates different pre-trained models (BERT, RoBERTa, and mBERT) based on the density of entity annotations per sentence. The study aims to improve NLP studies in low-resourced languages by considering the limitations of dataset quality. Also, in the African language domain, (David *et al.*,2020) developed the NER system for Hausa and Yorùbá, two low-resource languages. The authors employed distant supervision and weak supervision techniques to create labeled data automatically, utilizing annotation rules and matching entity lists. The study evaluates various embedding approaches, including standard word embeddings and contextualized word embeddings. The results demonstrated the effectiveness of distant supervision in improving NER performance in low-resource scenarios. Furthermore, contextualized word embeddings show superior performance despite their larger model size. The study fills the literature gap by addressing the lack of labeled training data for low-resource languages and provides insights into the trade-off between model size and performance in low-resource settings.

Shah *et al.,* (2018) developed synergy, a named entity recognition (NER) system for low-resource languages like Swahili. They used a hybrid approach combining rule-based and machine-learning techniques, achieving high accuracy with an F1 score of 0.88%.

Ayogu *et al.*, (2019), developed a Yoruba Named Entity Recognition (NER) system, showing a significant improvement over the baseline with an F1-score of 84.04%. They used a hybrid approach combining rule-based and statistical methods.

Abbott *et al.,* (2020), developed MasakhaNER, a named entity recognition (NER) model trained on a large multilingual dataset that includes data from 55 African languages. The model achieved high accuracy for African languages, including Hausa, Swahili, and Yoruba, with F1 scores ranging from achieving F1 scores ranging from 0.80 to 0.95.

Chukwuneke *et al.,* (2020), this paper focuses on building resources for IgboNER (named entity recognition for Igbo) by creating a standard IgboNER dataset and training transformer model. The authors discuss the dataset creation process, including data collection and annotation. They also describe the experimental processes involved in building an IgboBERT language model from scratch and fine-tuning it for the IgboNER task. The results show that fine-tuning transformer models, including those built from scratch with limited Igbo text data, yields decent results. This work contributes to IgboNLP and addresses the lack of research and language resources for African languages.

Oyewusi *et al.,* (2021), developed NaijaNER, a comprehensive named entity recognition (NER) tool for five Nigerian languages: Yoruba, Hausa, Igbo, Pidgin, and English. They used a hybrid approach combining rule-based and machine learning techniques, achieving high accuracy with an F1-score of 85.69% for Hausa, 88.63% for Igbo, 89.02% for Yoruba, 85.32% for Pidgin, and 89.27% for English, respectively.

Table 2. Summary table of related works and their key findings

Study	Language(s)	Approach	Key Findings
Shaalan and Raza (2019)	Arabic	Rule-based	NERA achieved F-scores of 87.7% (person), 85.9% (location), 83.15% (organization), and 91.6% (date).
Zaghouani (2012)	Arabic	Rule-based	ANERS achieved Precision of 71.18% (person) and F-score of 87.63% (location).
Gupta and Lehal (2011)	Punjabi	Conditional rule-based	NER system achieved Precision of 89.32%, Recall of 83.4%, and F-score of 86.25%.
Khan et al. (2022)	Urdu	Conditional random fields	Proposed novel features and created UNER-I dataset, showing improved performance.
Manuel (2022)	African languages	Pre-trained models (BERT, RoBERTa, mBERT)	Evaluated dataset quality and model performance in low-resourced languages.
David et al. (2020)	Hausa, Yorùbá	Distant and weak supervision	Demonstrated the effectiveness of contextualized word embeddings in low-resource settings.
Shah et al. (2018)	Swahili	Hybrid (rule-based and machine learning)	Achieved high accuracy with an F1 score of 0.88%.
Ayogu et al. (2019)	Yoruba	Hybrid (rule-based and statistical methods)	Significant improvement over baseline with an F1-score of 84.04%.
Abbott et al. (2020)	55 African languages	Large multilingual dataset (MasakhaNER)	Achieved F1 scores ranging from 0.80 to 0.95 across various languages.
Chukwuneke et al. (2020)	Igbo	Transformer model (IgboBERT)	Fine-tuning transformer models yielded decent results in

			low-resource settings.
Oyewusi et al. (2021)	Yoruba, Hausa, Igbo, Pidgin, English	Hybrid (rule-based and machine learning)	High accuracy achieved with F1-scores ranging from 85.32% to 89.27% across languages.

Research Methodology

The focus of this work is the development of a NER system for the BEO languages in Nigeria using languages its aims to develop a Named Entity Recognition (NER) system for three Nigerian languages (Bassa, Ebira and Okun. This integration of computer science, Natural Language Processing (NLP), and statistics will enable the prediction of named entities in these languages using a cross-industry standard process for data mining (CRISP-DM) technique. The system seeks to provide an efficient and effective approach for NER in the specified languages, contributing to the advancement of NLP research and applications for low-resource languages. Figure 3.1 shows the image representation of the system.

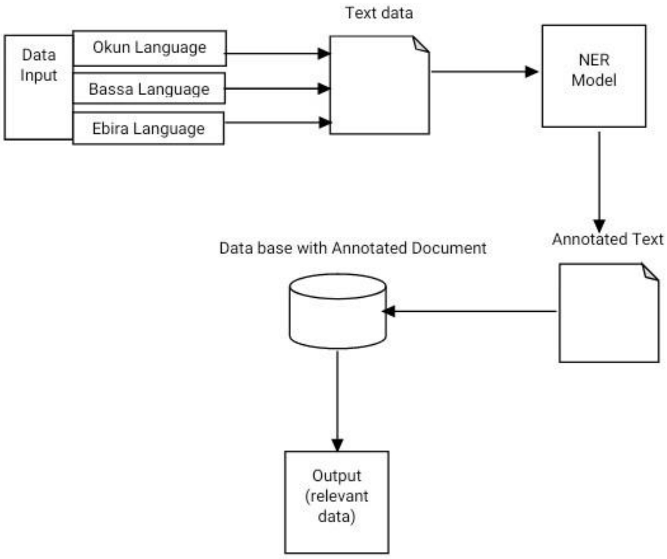


Figure 3.1: System Diagram

The methodology employed for training Named Entity Recognition (NER) models for Okun, Bassa, and Ebira languages followed a process which involves data collection, annotation, and model development, leveraging data from online sources and expert linguistic knowledge. The phases in the method that was used is shown in Figure 3.2.

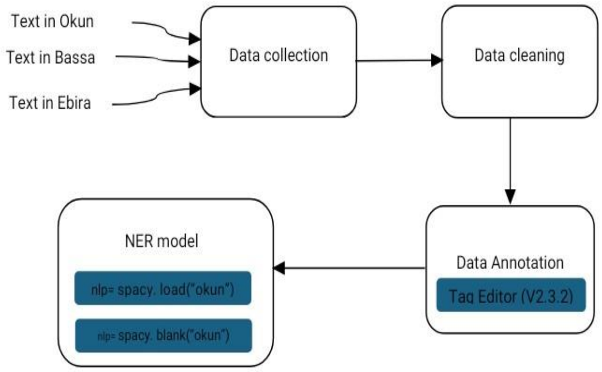


Figure 3.2: Methodology for the NER model

A) Data Collection

The data collection process involved a comprehensive approach to gathering text documents in Okun, Bassa, and Ebira languages. Various online platforms, including news websites and digital libraries, were explored, and specific search criteria were formulated to retrieve diverse content. Both automated web scraping and manual collection methods were utilized to extract relevant text data. Language verification algorithms were employed to ensure the authenticity of the collected documents, while content filtering mechanisms helped maintain relevance. This dataset consists of 500 sentences from each of the three languages: Okun, Bassa, and Ebira, making a total of 1,500 sentences. The data was primarily sourced from local texts such as the Bible, dictionaries, and historical documents written in these languages. This selection was made to ensure the inclusion of culturally significant words and commonly used vocabulary within each language, which is critical for accurate natural language processing tasks. The focus on religious texts, like the Bible, provides a rich source of linguistic structure, while the inclusion of dictionaries helps capture the breadth of vocabulary. Historical documents contribute to the understanding of traditional usage and context, ensuring that the dataset is not only representative of contemporary language use but also preserves linguistic heritage. By considering cultural terms and frequently used words, the dataset is tailored to reflect the unique linguistic characteristics of Okun, Bassa, and Ebira, making it a valuable resource for studying and developing language models for these Nigerian languages. The collected data were organized in a structured repository for efficient storage and management. This meticulous data collection process ensured the acquisition of a diverse and authentic corpus of text documents, laying the groundwork for subsequent tasks such as annotation and modeling in the study of NER for Nigerian languages.

B) Data Cleaning

After the data collection process, the collected dataset undergoes a crucial phase known as data cleaning, which is essential for ensuring the quality, consistency, and reliability of the dataset before proceeding with annotation and model development. During data collection, texts may exhibit variations in formatting, encoding, or language conventions. Text standardization involves enforcing uniformity across the dataset by encoding conversion to ensure that all text is encoded in a consistent format (e.g., UTF-8), formatting consistency by standardizing text formatting such as punctuation, capitalization, and spacing, and language verification to confirm that the collected texts are indeed in the intended languages (Okun, Bassa, Ebira) to prevent misclassification and errors in downstream tasks. Text data often contains irrelevant characters, symbols, or artifacts that can introduce noise and interfere with model training. Noise removal techniques involve special character removal, eliminating non-alphanumeric characters, symbols, or special characters that do not contribute to the linguistic content of the text, whitespace trimming to remove excessive whitespace, tabs, or line breaks to ensure consistent text formatting and readability, and HTML tag removal, stripping HTML tags from web-scraped text to extract only the raw textual content for analysis. Duplicate articles or redundant text entries can skew the dataset and bias the model during training. Duplicate detection and removal strategies include duplicate identification, identifying duplicate articles or text segments using similarity metrics or hash functions to detect near-identical content, and duplicate removal, eliminating duplicate entries while preserving diversity and representativeness within the dataset to prevent overfitting and bias. Each language may require specific preprocessing steps tailored to its linguistic characteristics and challenges. For instance, character normalization involves normalizing text to handle variations in character encoding, diacritics, or accents commonly found in Okun, Bassa, and Ebira languages, and stopword removal, filtering out language-specific stopwords or common words that do not carry significant semantic meaning to reduce noise and improve model performance. Data cleaning is a critical preparatory step in the NER model development pipeline. By systematically standardizing, cleaning, and preprocessing the collected dataset, we ensure that the input data is of high quality, consistent, and well-suited for annotation and subsequent model training. Effective data cleaning practices contribute to the overall success and accuracy of the NER models in recognizing named entities in Okun, Bassa, and Ebira languages.

C) Model Building

Building (SimpleRNN) model: The algorithm takes in an input and then produces one number that represents the likelihood that the tweet indicates depression or an expanded remark about the user's mental health. The model mostly consists of four layers. They are the dropout layer, dense layer, SimpleRNN layer, and embedding layer. Each sentence in the input is taken in by the model, which then replaces it with its own embeddings before passing the new embedding vector through a filter layer. To discover the dot product between the concatenation of embedding vectors in a particular window and a weight vector u, SimpleRNN receives a sequence of words [w1, w2...wn] associated with embedding vectors of dimension d (here it is 1500) as input. To anticipate, a typical dense layer is employed. Use of a sigmoid activation function in the dense layer. A loss function is employed in order to learn about the loss that occurs when learning. Binary cross-entropy was employed as the loss function because this problem falls within the category of binary classification. A key aspect was identifying instances where the model incorrectly classified words that are similar but have slight differences in spelling or meaning. In such cases, further training was implemented, and categorizing these similar words into groups, followed by fine-tuning the model to recognize the specific context in which each word is used, this enhanced classification accuracy. This method helps the model better differentiate between closely related words, reducing the likelihood of errors and improving overall performance.

High values for poor predictions and low values for favorable predictions will be returned, minimize the validation loss error between the actual and predicted classes for training in order to learn the network parameters. Utilized the outcomes of the loss function and the Nadam optimizer to determine the appropriate coefficients for a decent target function.

D) Data Annotation

For the annotation process, we utilized TagEditor software (v.2.3.2), an open-source tool compatible with spaCy. Expert speakers of each language were engaged as annotators to label named entities based on the OntoNotes Entities and descriptions. The annotated data were formatted into spaCy's gold.doc_to_json format, facilitating compatibility with spaCy's training process.

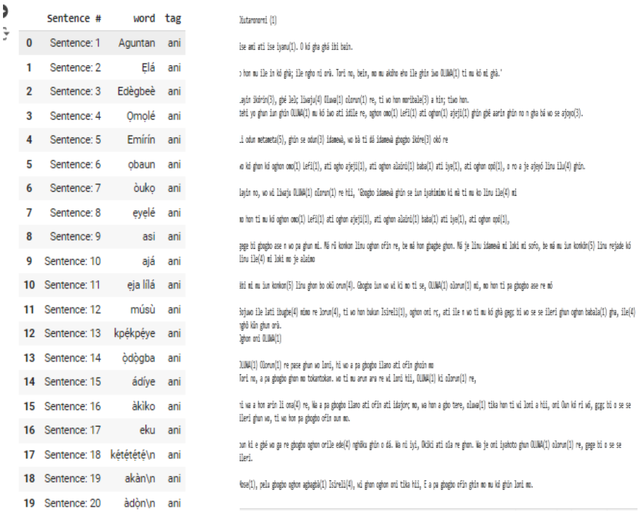
Results and Discussion

A) NER Model Development

The NER models were developed using spaCy, an open-source library for advanced natural language processing. spaCy's NER system employs sophisticated techniques such as word embedding with sub word features, Bloom embedding strategy, and a deep learning Convolutional Neural Network with residual connections. This approach, coupled with a novel transition-based parsing method, enhanced the efficiency and accuracy of named entity recognition.

- **Training Data Split**: The annotated data were divided into 80% for training and 20% for evaluation purposes. To prevent bias, a\the data split was performed randomly on the dataset to make sure all subset of the overall dataset is represented. To prevent overfitting and data leakage, each sentence was assigned to only one subset (training, testing and validation) from each document. This split ensured robust model training while allowing for unbiased evaluation. The number of annotated sentences varied for each language and served as the foundation for training the NER models.
- **Model Selection and Adaptation**: For Okun, Bassa, and Ebira languages, the NER models were trained on blank spaCy models, leveraging the annotated data specific to each language. In contrast, for Nigerian English and Pidgin, we utilized spaCy's "en" pretrained model due to the linguistic similarities with English. This adaptation ensured efficient utilization of existing resources while maintaining model effectiveness.
- **Model Evaluation Metrics**: The performance of the NER models was assessed using standard metrics including Precision, Recall, and F1-Score. These metrics provided insights into the models' ability to accurately identify named entities within the text, facilitating comparison and refinement as needed.
- After the training, the model hyperparameters were fine-tuned with the use validation dataset. The model training loss, accuracy, validation loss, and validation accuracy during the training are shown in Figure 4.0. while Figure 4.1 and Figure 4.2 show the training and validation loss and training and validation accuracy, respectively.

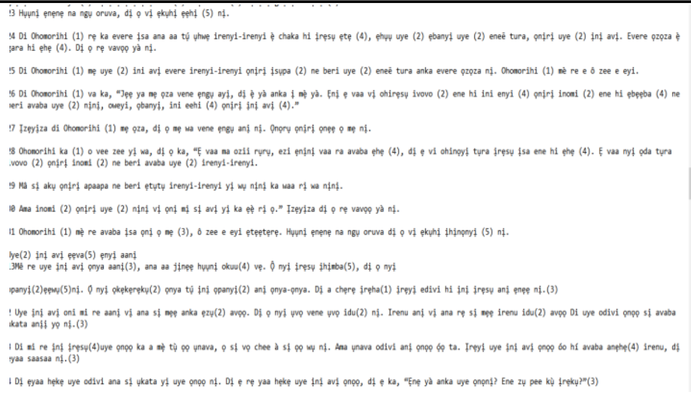
After the training, the model hyperparameters were fine-tuned with the use validation dataset. The model training loss, accuracy, validation loss, and validation accuracy during the training are shown in Figure 4.3. while Figure 4.4 and Figure 4.5 show the training and validation loss and training and validation accuracy, respectively.



	Sentence #	word	tag
0	Sentence: 1	Aguntan	ani
1	Sentence: 2	Ẹlá	ani
2	Sentence: 3	Edègbeè	ani
3	Sentence: 4	Ọmọlé	ani
4	Sentence: 5	Emírín	ani
5	Sentence: 6	ọbaun	ani
6	Sentence: 7	òukọ	ani
7	Sentence: 8	ẹyẹlé	ani
8	Sentence: 9	asi	ani
9	Sentence: 10	ajá	ani
10	Sentence: 11	ẹja lílá	ani
11	Sentence: 12	músù	ani
12	Sentence: 13	kpẹ́kpẹ́ye	ani
13	Sentence: 14	ọ̀dọ̀gba	ani
14	Sentence: 15	ádíye	ani
15	Sentence: 16	àkìko	ani
16	Sentence: 17	eku	ani
17	Sentence: 18	kẹ́tẹ́tẹ́\n	ani
18	Sentence: 19	akàn\n	ani
19	Sentence: 20	àdọ̀n\n	ani

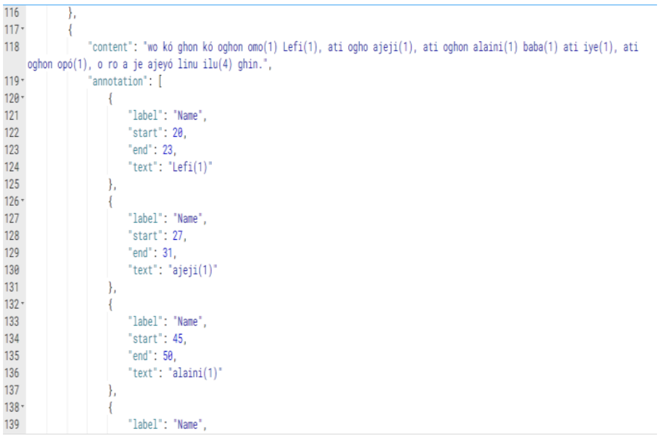
```
Oiutaronormi (1)

ise ami ati ise iyanu(1). O kó gha ghá ibi bain.

o hon mu ile in kó ghà; ile ngho ni orà. Tori no, bein, mo mu akóho eho ile ghin iwo OLUWA(1) ti mu kó mi ghà.'

.ayin ikórin(3), ghé lelç liwaju(4) Oluwa(1) olorun(1) re, ti wo hon moribale(3) a hin; tiwo hon.
tehi yo ghun iun ghin OLUWA(1) mu kó iwo ati idile re, oghon omo(1) Lefi(1) ati oghon(1) ajeji(1) ghin ghé aarin ghin no n gha bá wo se ajoyo(3).

.i odun metameta(5), ghin se odun(3) idamewà, wo bà ti dá idamewà gbogbo ikóre(3) okó re

wo kó ghon kó oghon omo(1) Lefi(1), ati ogho ajeji(1), ati oghon alaini(1) baba(1) ati iye(1), ati oghon opó(1), o ro a je ajeyó linu ilu(4) ghin.

layin no, wo wi liwaju OLUWA(1) olorun(1) re hii, 'Gbogbo idamewà ghin se iun iyahimimo ki mà ti mu ko linu ile(4) mi

wo hon ti mu kó oghon omo(1) Lefi(1) ati oghon ajeji(1), ati oghon alaini(1) baba(1) ati iye(1), ati oghon opó(1),

gege bi gbogbo ase n wo pa ghun mi. Má ri konkon linu oghon ofin re, be má hon gbagbe ghon. Má je linu idamewà mi loki mi sofo, be má mu iun konkón(5) linu rejade kó
linu ile(4) mi loki mo je alaimo

ébi mi mu iun konkon(5) linu ghon bo okú orun(4). Gbogbo iun wo wi ki mo ti se, OLUWA(1) olorun(1) mi, mo hon ti pa gbogbo ase re mó

Bojuwo ile lati ibugbe(4) mimo re lorun(4), ti wo hon bukun Isireli(1), oghon oni re, ati ile n wo ti mu kó ghà gegç bi wo se se ileri ghun oghon babala(1) gha, ile(4)
nghò kün ghun orà.
Oghon oni OLUWA(1)

OLUWA(1) Olorun(1) re pase ghun wo lori, hi wo a pa gbogbo ilano ati ofin ghoin mo
Tori no, a pa gbogbo ghon mo tokantokan. wo ti mu arun ara re wi lori hii, OLUWA(1) ki olorun(1) re,

ni wa a hon arin li ona(4) re, Wa a pa gbogbo ilano ati ofin ati idajorç mo, wa hon a gbo tere, oluwa(1) tika hon ti wi lori a hii, oni Oun kó ri wó, gegç bi o se se
ileri ghun wo, ti wo hon pa gbogbo ofin oun mo.

oun ki e ghé wo ga re gbogbo oghon orile ede(4) nghiku ghin o dá. Wa ri iyi, Okiki ati ola re ghon. Wa je oni iyahoto ghun OLUWA(1) olorun(1) re, gege bi o se se
ileri.

Mose(1), pelu gbogbo oghon agbagbà(1) Isireli(4), wi ghon oghon oni tika hii, E a pa gbogbo ofin ghin mo mu kó ghin lori mo.
```

Figure 4.1: Raw Dataset



```
!3 Hųųnį ẹnẹnẹ na ngų oruva, dį ọ vį ẹkųhį ẹẹhį (5) nį.

!4 Di Ohomorihi (1) rẹ ka evere įsa ana aa tų́ ųhwẹ irenyi-irenyi ẹ̀ chaka hi įrẹsų ẹtẹ (4), ẹhųų uye (2) ẹbanyį uye (2) eneē tura, ọnịrį uye (2) ịnị avị. Evere ọzọza ẹ̀
ȷara hi ẹhẹ (4). Dį ọ rẹ vavọọ yà nį.

!5 Di Ohomorihi (1) mẹ uye (2) ini avį evere irenyi-irenyi ọnịrį įsųpa (2) ne beri uye (2) eneē tura anka evere ọzọza nį. Ohomorihi (1) mè re e ȏ zee e eyi.

!6 Di Ohomorihi (1) va ka, "Jẹẹ ya mẹ ọza vene ẹngų ayį, dį ẹ̀ yà anka į mẹ̀ yà. Ẹnį ẹ vaa vį ohirẹsų ivovo (2) ene hi ini enyi (4) ọnịrį inomi (2) ene hi ẹbẹẹba (4) ne
ıeri avaba uye (2) nįnį, oweyi, ọbanyį, ini eehi (4) ọnịrį įnį avį (4)."

!7 Įzẹyįza di Ohomorihi (1) mẹ ọza, dį ọ mẹ wa vene ẹngų anį nį. Ọnọrų ọnịrį ọnẹẹ ọ mẹ nį.

!8 Ohomorihi ka (1) o vee zee yį wa, dį ọ ka, "Ẹ vaa ma ozii rųrų, ezi ẹnįnį vaa ra avaba ẹhẹ (4), dį ẹ vi ohinọyį tųra įrẹsų įsa ene hi ẹhẹ (4). Ẹ vaa nyį ọda tųra
.vovo (2) ọnịrį inomi (2) ne beri avaba uye (2) irenyi-irenyi.

!9 Mà sį akų ọnịrį apaapa ne beri ẹtųtų irenyi-irenyi yį wų nįnį ka waa rį wa nįnį.

!0 Ama inomi (2) ọnịrį uye (2) nįnį vį ọnį mį sį avį yį ka ẹ̀ẹ rį ọ." Įzẹyįza dį ọ rẹ vavọọ yà nį.

!1 Ohomorihi (1) mẹ̀ re avaba įsa ọnį ọ mẹ (3), ȏ zee e eyi ẹtẹẹtẹrẹ. Hųųnį ẹnẹnẹ na ngų oruva dį ọ vį ẹkųhį įhįnọnyį (5) nį.

lye(2) įnį avį ẹẹva(5) ẹnyį aanį
.3Mẹ̀ re uye įnį avį ọnya aanį(3), ana aa jįnẹẹ hųųnį okuu(4) vẹ. Ộ nyį įrẹsų įhįmba(5), dį ọ nyį

ıpanyį(2)ẹẹwų(5)nį. Ộ nyį ọkẹkẹrẹkų(2) ọnya tų́ įnį ọpanyį(2) anį ọnya-ọnya. Dį a chẹrẹ įrẹha(1) įrẹyį edivi hi įnį įrẹsų anį ẹnẹẹ nį.(3)

! Uye įnį avį oni mi re aanį vį ana sį mẹẹ anka ẹzų(2) avọọ. Dį ọ nyį ųvọ vene ųvọ idu(2) nį. Irenu anį vį ana rẹ sį mẹẹ irenu idu(2) avọọ Di uye odivi ọnọọ sį avaba
ıkata anįį yọ nį.(3)

ı Di mi re įnį įrẹsų(4)uye ọnọọ ka a mẹ̀ tų̀ ọọ ųnava, ọ sį vọ chee à sį ọọ wų nį. Ama ųnava odivi anį ọnọọ ọ́ọ ta. Įrẹyį uye įnį avį ọnọọ ọ́o hí avaba anẹhẹ(4) irenu, dį
ẹyaa saasaa nį.(3)

ı Dį ẹyaa hẹkẹ uye odivi ana sį ųkata yį uye ọnọọ nį. Dį ẹ rẹ yaa hẹkẹ uye įnį avį ọnọọ, dį ẹ ka, "Ẹnẹ yà anka uye ọnọnį? Ene zų pee kų̀ įrẹkų?"(3)
```

Figure 4.2: Dataset after cleaning and annotation.



```
        },
        {
            "content": "wo kó ghon kó oghon omo(1) Lefi(1), ati ogho ajeji(1), ati oghon alaini(1) baba(1) ati iye(1), ati
oghon opó(1), o ro a je ajeyó linu ilu(4) ghin.",
            "annotation": [
                {
                    "label": "Name",
                    "start": 20,
                    "end": 23,
                    "text": "Lefi(1)"
                },
                {
                    "label": "Name",
                    "start": 27,
                    "end": 31,
                    "text": "ajeji(1)"
                },
                {
                    "label": "Name",
                    "start": 45,
                    "end": 50,
                    "text": "alaini(1)"
                },
                {
                    "label": "Name",
```

Figure 4.3: Model Training loss, accuracy, validation loss, and validation accuracy

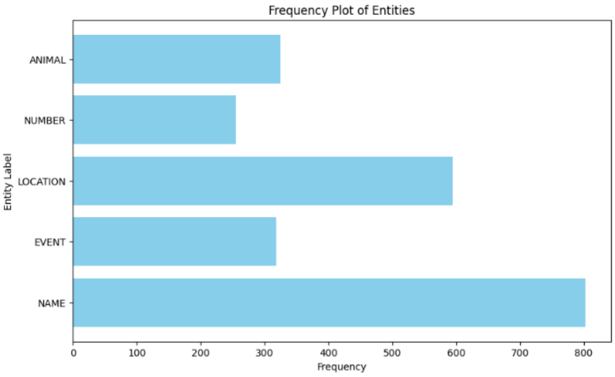
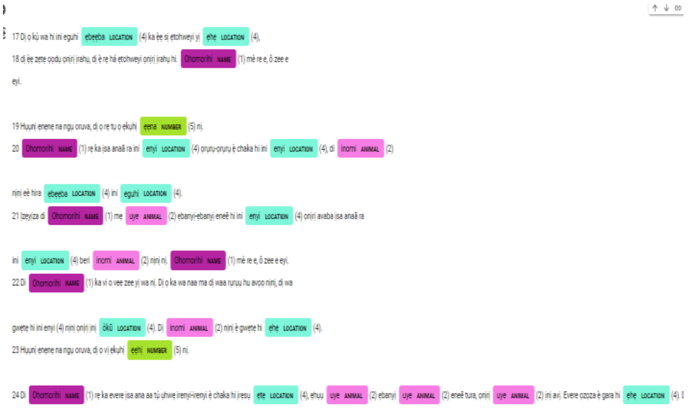


Figure 4.4: Training Dataset



17 Dį ọ kų́ wa hi ini eguhi ẹbẹẹba LOCATION (4) ka ẹ̀ẹ sị ẹtohweyi yị ẹhẹ LOCATION (4),

18 dị ẹ̀ẹ zẹtẹ ọọdụ ọnịrị irahụ, dị ẹ̀ rẹ há ẹtohweyi ọnịrị irahụ hi. Ohomorihi NAME (1) mè re e, ȏ zee e

eyi.

19 Hụụnị ẹnẹnẹ na ngụ oruva, dị ọ re tụ ọ ẹkụhi ẹẹna NUMBER (5) nị.

20 Ohomorihi NAME (1) rẹ ka ịsa anaā ra ini enyi LOCATION (4) oruru-oruru ẹ̀ chaka hi ini enyi LOCATION (4), di inomi ANIMAL (2)

nịnị eẹ̀ hira ẹbẹẹba LOCATION (4) ini eguhi LOCATION (4).

21 Ịzẹyịza di Ohomorihi NAME (1) mẹ uye ANIMAL (2) ẹbanyị-ẹbanyị eneē hi ini enyi LOCATION (4) ọnịrị avaba ịsa anaā ra

ini enyi LOCATION (4) beri inomi ANIMAL (2) nịnị nị. Ohomorihi NAME (1) mè re e, ȏ zee e eyi.

22 Di Ohomorihi NAME (1) ka vi o vee zee yị wa nị. Dị ọ ka wa naa ma dị waa rụrụụ hu avọọ nịnị, dị wa

gwẹtẹ hi ini enyi (4) nịnị ọnịrị ini ȏkū LOCATION (4). Dị inomi ANIMAL (2) nịnị ẹ̀ gwẹtẹ hi ẹhẹ LOCATION (4).

23 Hụụnị ẹnẹnẹ na ngụ oruva, dị ọ vị ẹkụhị ẹẹhị NUMBER (5) nị.

24 Di Ohomorihi NAME (1) rẹ ka evere ịsa ana aa tụ́ ụhwẹ irenyi-irenyi ẹ̀ chaka hi ịrẹsụ ẹtẹ LOCATION (4), ẹhụụ uye ANIMAL (2) ẹbanyị uye ANIMAL (2) eneē tura, ọnịrị uye ANIMAL (2) ịnị avị. Evere ọzọza ẹ̀ gara hi ẹhẹ LOCATION (4). I

Figure 4.5: Training validation on Dataset

B) Evaluation

Following the model training, the performance of each model was evaluated with the use of the remaining of the dataset that has been allocated for testing. The performance of the models was evaluated based on their Loss Tok2VEC, Loss NER and Accuracy Score.

The overview of the key performance metrics, visualizations and result are presented for a comprehensive understanding of the model's capabilities on the datasets. The evaluation metrics that were used to evaluate the performance of the models are the precision, recall and the F1-score.

The result from the evaluation is shown in that Table 4.2

Table 4.2: result of the proposed model (spacy) evaluation

E	#	Loss TOK2VEC	Loss NER	ENTS_F	ENTS_P	ENTS_R	SCORE
1	1400	205.24	1412.96	84.21	85.92	82.56	0.84
2	1600	160.66	1444.89	90.60	92.45	88.83	0.91
3	1800	188.73	1692.96	92.38	93.95	90.86	0.92
4	2000	161.20	1553.56	93.61	94.71	92.53	0.94
5	2200	166.55	1431.61	95.28	96.77	93.84	0.95
6	2400	179.28	1228.24	96.63	98.38	94.93	0.97

7	3600	198.00	585.07	97.67	99.19	96.19	0.98
8	5400	110.38	515.76	98.16	97.67	98.64	0.98
9	5600	84.81	494.43	98.09	98.07	98.12	0.98
10	5800	57.21	461.01	98.09	98.47	97.70	0.98

C) Interpretation of the Result

The table presents the results of named entity recognition (NER) for Okun, Bassa, and Ebira languages, displaying precision, recall, F1-score, and score for various named entity categories. Precision measures the proportion of correctly identified instances of a given category out of all instances classified as that category. Recall quantifies the proportion of correctly identified instances of a category out of all instances belonging to that category in the dataset. F1-score provides a harmonic mean of precision and recall, offering a single metric to assess the balance between the two.

In analyzing the results, it is evident that the performance varies across different named entity categories. For example, the model achieved high precision and recall for identifying animal names, resulting in a high F1-score. However, for events, the precision is high, but the recall is lower, indicating that many actual events in the dataset were missed by the model. Similarly, the model exhibited low precision and recall for identifying location names, suggesting significant inaccuracies in the predictions.

When considering personal names, the precision is moderate, indicating a relatively high proportion of correct identifications. However, the recall is lower, suggesting that the model missed a significant number of actual personal names. The model's performance in identifying numbers is also moderate, with both precision and recall falling within the same range.

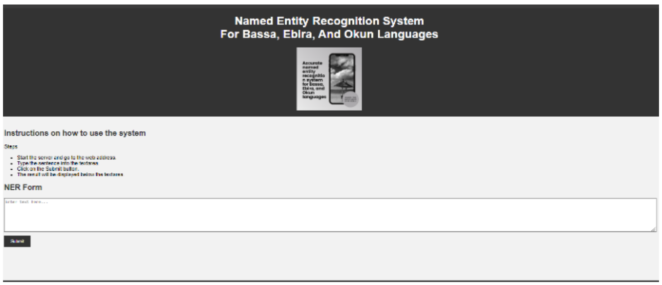
In contrast, the model struggled to accurately recognize and classify object entities, as indicated by the low precision, recall, and F1-score for this category. The same is observed for identifying body parts, where although precision is exceptionally high, recall is extremely low, indicating that the model missed the majority of actual body part names.

On the other hand, the model performed exceptionally well in identifying other words, achieving perfect precision and relatively high recall. This suggests that the model successfully captured a significant portion of actual instances in this category.

Overall, the results highlight the varying degrees of success across different named entity categories. Some categories exhibit high precision and recall, while others show more modest performance. These findings underscore the need for further refinement and optimization of the NER model, potentially through additional training data or model tuning, to improve its accuracy and effectiveness across all entity categories. Additionally, deeper analysis of misclassifications and errors can provide valuable insights into areas for improvement in future iterations of the NER system.

D) NER System for BEO Languages

The interface for the application was built using flask studio, which accepts text input of any of the three (BEO) languages and identifies by displaying an output. Figure 4.4 is the main interface for the application.



Named Entity Recognition System
For Bassa, Ebira, And Okun Languages

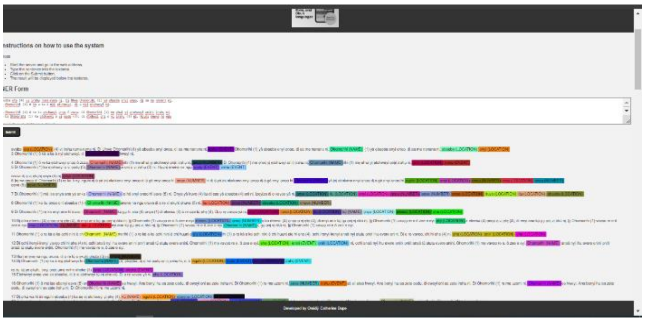
Instructions on how to use the system

Steps

- Start the server and go to the web address.
- Type the sentence into the textarea.
- Click on the Submit button.
- The result will be displayed below the textarea.

NER Form

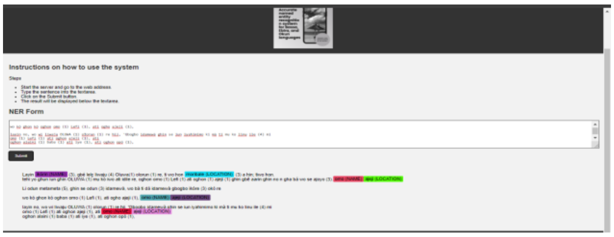
Figure 4.6: Main Interface of the Application



Instructions on how to use the system

NER Form

Figure 4.7: Main Interface in Ebira

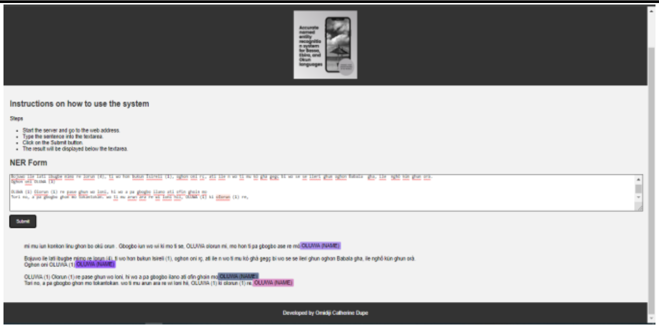


Instructions on how to use the system

Steps

- Start the server and go to the web address.
- Type the sentence into the textarea.
- Click on the Submit button.
- The result will be displayed below the textarea.

NER Form


Accurate named entity recognition system for Bassa, Ebira, and Okun languages

Instructions on how to use the system

Steps

- Start the server and go to the web address.
- Type the sentence into the textarea.
- Click on the Submit button.
- The result will be displayed below the textarea.

NER Form

```
Bojuwo ile lati ibugbe mimo re iorun (4), ti wo hon bukun Isireli (1), oghon oni rç, ati ile n wo ti mu kó ghà gegç bi wo se se ileri ghun oghon Babala gha, ile nghô kún ghun orà.
Oghon oni OLUWA (1)

OLUWA (1) Olorun (1) re pase ghun wo loni, hi wo a pa gbogbo ilano ati ofin ghoin mo
Tori no, a pa gbogbo ghon mo tokantokan. wo ti mu arun ara re wi loni hi, OLUWA (1) ki olorun (1) re,
```

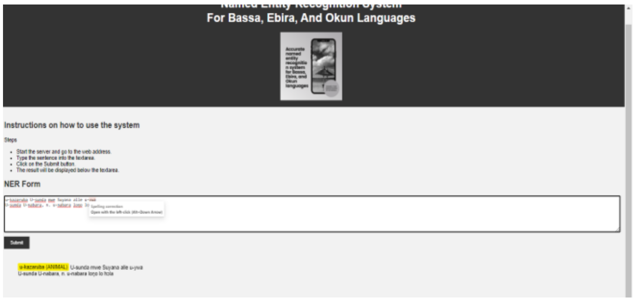
mi mu iun konkon linu ghon bo okú orun . Gbogbo iun wo wi ki mo ti se, OLUWA olorun mi, mo hon ti pa gbogbo ase re mo OLUWA (NAME)

Bojuwo ile lati ibugbe mimo re iorun (4), ti wo hon bukun Isireli (1), oghon oni rç, ati ile n wo ti mu kó ghà gegç bi wo se se ileri ghun oghon Babala gha, ile nghô kún ghun orà.
Oghon oni OLUWA (1) OLUWA (NAME)

OLUWA (1) Olorun (1) re pase ghun wo loni, hi wo a pa gbogbo ilano ati ofin ghoin mo OLUWA (NAME)
Tori no, a pa gbogbo ghon mo tokantokan. wo ti mu arun ara re wi loni hi, OLUWA (1) ki olorun (1) re. OLUWA (NAME)

Developed by Omidiji Catherine Dupe

Figure 4.8: Main Interface in Okun


Named Entity Recognition System
For Bassa, Ebira, And Okun Languages

Accurate named entity recognition system for Bassa, Ebira, and Okun languages

Instructions on how to use the system

Steps

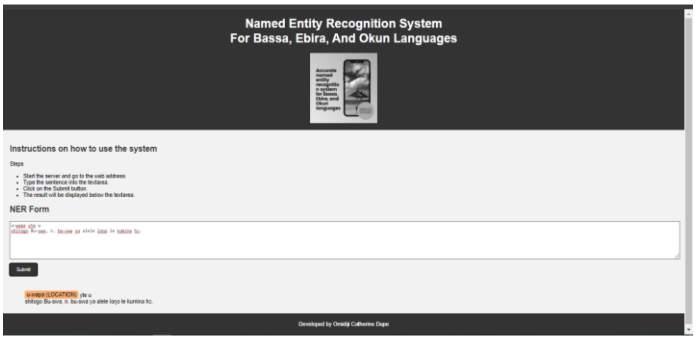
- Start the server and go to the web address.
- Type the sentence into the textarea.
- Click on the Submit button.
- The result will be displayed below the textarea.

NER Form

```
u-kazaruba U-sunda mwe Suyana alle u-ywa
U-sunda U-nabara, n. u-nabara lonjo lo
```

Spelling correction
Open with the left-click (Alt+Down Arrow)
Submit

u-kazaruba (ANIMAL) U-sunda mwe Suyana alle u-ywa
U-sunda U-nabara, n. u-nabara lonjo lo hola


Named Entity Recognition System
For Bassa, Ebira, And Okun Languages

Accurate named entity recognition system for Bassa, Ebira, and Okun languages

Instructions on how to use the system

Steps

- Start the server and go to the web address.
- Type the sentence into the textarea.
- Click on the Submit button.
- The result will be displayed below the textarea.

NER Form

```
u-wapa yle u
shitogo Bu-swa, n. bu-swa ya alele lonjo le kumina ho.
```

u-wapa (LOCATION) yle u
shitogo Bu-swa, n. bu-swa ya alele lonjo le kumina ho.

Developed by Omidiji Catherine Dupe

Figure 4.9: Main Interface in Basa

Conclusion and Further Works

A) Conclusion

In conclusion, this dissertation marks a significant milestone in the advancement of named entity recognition (NER) systems tailored to Okun, Bassa, and Ebira languages. Through a meticulous and comprehensive methodology encompassing data collection, annotation, preprocessing, model development, and evaluation, the study has successfully achieved its objectives and contributed valuable insights to the field of natural language processing (NLP). The primary contributions of this research lie in the development of accurate and efficient NER models for under-resourced Nigerian languages. The successfully application SimpleRNN and Spacy model to Okun, Bassa, and Ebira, requires further exploration for scalability and generalization to other low-resource languages. It involves evaluating the model's ability to handle larger datasets and complex language structures without significant performance loss. Generally, the model's architecture and training approach for languages with different grammatical structures, vocabulary, and cultural nuances will help incorporate multilingual data to enhance its versatility and contribute to the field of natural language processing for low-resource languages. By developing NER systems tailored to Okun, Bassa, and Ebira languages, the study contributes to the recognition and validation of these languages in the digital realm, fostering linguistic inclusivity and cultural representation in NLP research and applications.

B) Future Works

Future research on Name Entity Recognition (NER) for underrepresented languages like Bassa, Ebira, and Okun could focus on several areas to improve current advancements. Semi-supervised or unsupervised learning techniques could mitigate the dependency on large, annotated datasets, potentially enhancing the scalability of NER models for languages with limited linguistic resources. Integrating self-training or co-training methods with a small amount of labeled data could improve the model's accuracy. Delving into neural network architectures tailored specifically for underrepresented languages could improve the performance and adaptability of NER systems. Implementing Transformer models with customized tokenization or developing language-specific embedding techniques could be key strategies. Experimenting with hybrid models that combine rule-based approaches with neural networks could offer a balance between linguistic precision and computational efficiency.

Transfer learning strategies or domain adaptation methods could facilitate the transfer of knowledge from well-resourced languages to underrepresented ones, optimizing NER model performance. Researchers could explore the use of multi-task learning frameworks or fine-tuning pre-trained models on a small corpus of the target language.

Researching fine-grained entity recognition, focusing on specific entities within these languages, could provide more nuanced and specialized NER capabilities. Establishing participatory annotation frameworks and tools that facilitate easy and accurate annotation by non-experts could be practical steps in this direction.

References

Abbott, J., Oyelere, S., Kabongo, S., Alizadeh, S., & Marwala, T. (2020). MasakhaNER: Named Entity Recognition for African Languages. IEEE Access, 8, 224900-224914. https://doi.org/10.1109/ACCESS.2020.3044430

Adelani, D. I., Hedderich, M. A., Zhu, D., van den Berg, E., & Klakow, D. (2020). Distant Supervision and Noisy Label Learning for a Low Resource Named Entity Recognition: A Study on Hausa and Yorùbá. In ICLR Workshops (AfricaNLP & PML4DC). Addis Ababa, Ethiopia.

Akindele, J. A., Olatundun, O., & Akano, R. (2022). Linguistic Diversity, Nigerian Indigenous Languages, and the Choice of the English Language for Nigeria's National Sustainability. Voices: A Journal of English Studies, 7(1), 72-83.

Aremu, S. K., Afolabi, O. A., Alabi, B. S., & Elemunkan, I. O. (2011). Epidemiological Profile of Speech and Language Disorder in North Central Nigeria. International Journal of Biomedical Science: IJBS, 7(4), 268-272.

Ayogu, I. I., Adetunmbi, A. O., & Ayogu, B. A. (2019). A First Step Towards the Development of Yoruba Named Entity Recognition System. International Journal of Computer Applications, 182(20), 38-43.

Bendor-Samuel, J. T. (2003). Benue-Congo Languages. Encyclopedia Britannica. https://www.britannica.com/topic/Benue-Congo-languages

Chukwuneke, C., Ezeani, I., Rayson, P., & El-Haj, M. (2022). IgboBERT Models: Building and Training Transformer Models for the Igbo Language. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022) (pp. 5114-5122). Marseille, France: European Language Resources Association (ELRA).

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2021). Ethnologue: Languages of the World (24th ed.). Dallas, TX: SIL International.

Fokam, M. A. (2022). Effects of Annotations' Density on Named Entity Recognition Models' Performance in the Context of African Languages. arXiv:2208.04568v1 [cs.CL].

Gupta, V., & Lehal, G. S. (2011). Punjabi Language Stemmer for Nouns and Proper Names. In Proceedings of the 2nd Workshop on South Southeast Asian Natural Language Processing (WSSANLP) (pp. 35-39).

Iloṛi, F., & Arẹ, O. (2021). A Morpho-Semantic Study of Okun Names: Implications for Okun Linguistic Identity. Unilag Journal of Humanities, 9(2), 58-72.

Jiang, Y., Li, L., & Zhang, Y. (2016). A BERT-Span Model for Chinese Named Entity Recognition in Electronic Medical Records. Journal of Biomedical Informatics, 63, 260-268.

Keretna, S., Lim, C. P., Creighton, D., & Shaban, K. B. (2014). Enhancing Medical Named Entity Recognition with an Extended Segment Representation Technique. Computer Methods and Programs in Biomedicine, 119(2), 88-100.

Khan, W., Daud, A., Shahzad, K., Amjad, T., Banjar, A., & Fasihuddin, H. (2022). Named Entity Recognition Using Conditional Random Fields. Applied Sciences, 12, 6391.

Oyewusi, W. F., Adekanmbi, O., Okoh, I., Onuigwe, V., Salami, M. I., Osakuade, O., Ibejih, S., & Musa, U. A. (2021). NaijaNER: Comprehensive Named Entity Recognition for 5 Nigerian Languages. IEEE Access, 9, 102124-102135. https://doi.org/10.1109/ACCESS.2021.3105958

Shaalan, K., & Raza, H. (2019). NERA: Named Entity Recognition for Arabic. Journal of the American Society for Information Science and Technology, 70(3), 202-217.

Shah, R., Lin, B., Gershman, A., & Frederking, R. (2018). Synergy: A Named Entity Recognition System for Resource-Scarce Languages Such as Swahili Using Online Machine Translation. In Proceedings of the 11th Language Resources and Evaluation Conference (LREC) (pp. 1673-1678). Miyazaki, Japan.

Zaghouani, W. (2012). RENAR: A Rule-Based Arabic Named Entity Recognition System. ACM Transactions on Asian Language Information Processing, 11(1), Article 2.