NaijaNER : Comprehensive Named Entity Recognition for 5 Nigerian Languages.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, Usman Abdullahi Musa,

Data Science Nigeria, Lagos Nigeria
{wuraola, olubayo, ifeoma, sharon, mary, abdullahi, osakuade, vitus}@datasciencenigeria.ai

Abstract
Most of the common applications of Named Entity Recognition (NER) is on English and other highly available languages. In this work, we present our findings on Named Entity Recognition for 5 Nigerian Languages (Nigerian English, Nigerian Pidgin English, Igbo, Yoruba and Hausa). These languages are considered low-resourced, and very little openly available Natural Language Processing work has been done in most of them. In this work, individual NER models were trained and metrics recorded for each of the languages. We also worked on a combined model that can handle Named Entity Recognition (NER) for any of the five languages. The combined model works well for Named Entity Recognition(NER) on each of the languages and with better performance compared to individual NER models trained specifically on annotated data for the specific language. The aim of this work is to share our learning on how information extraction using Named Entity Recognition can be optimized for the listed Nigerian Languages for inclusion, ease of deployment in production and reusability of models. Models developed during this project are available on GitHub and an interactive web app here.

1 Introduction
Named entity recognition(NER) is sometimes referred to as entity chunking, extraction, or identification. It is a subtask of information extraction (Nadeau and Sekine, 2007) in Natural Language Processing. Named Entity recognition involves processing a text and identifying certain occurrences of words or expressions as belonging to particular categories of Named Entities (NE) (Mikheev et al., 1999). Named entity recognition (NER) is commonly dealt with as a sequence labelling job and has been one of the most successfully tackled NLP tasks (Li, 2021).

The phrase “Named Entity” was coined in 1996 at the 6th Message Understanding Conference (MUC) when the extraction of information from unstructured text became an important problem in the linguistic domain (Nadeau and Sekine, 2007).

The concept of Named entity recognition transcends language and the principle is the same. Natural Language Processing enables improved communication between human and machines. Some of the most important use cases for Named Entity Recognition involve large text data. These use cases range from recognizing Named Entities for document classification, entity linking, social media analysis and methods for training NER models. For example (Ritter et al., 2011) in their work, “Named Entity Recognition in Tweets” address the approaches and challenges related to the style of unintentional language being used in social media, and how the limitation of standard NLP tool degradation on tweets and their custom T-NER model gave better performance. This makes a case for context-specific training of NER models. (Shinyama and Sekine, 2004) in their work “Named Entity Discovery Using Comparable News Articles” studied the pattern of NER in news articles and the way common nouns are used. The techniques used involve classifying based on the time series distribution of the entities.

1.1 Named Entity Recognition for Low Resource Languages
There are about 7,000+ languages spoken across the world, but of these, only a small fraction are considered high-resource languages. Languages are called low-resource when lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications (Sciforce, 2019). Many Nigerian languages are considered a low resource, including 4 of the 5 languages listed in this project. Nigerian English is not considered a low resource because it is standard English but included in this project to allow for context-specific NER for enti...
ties peculiar to the Nigerian scenario.

Cross-lingual transfer learning is a technique which can compensate for the dearth of data, by transferring knowledge from high to low-resource languages. This has typically taken the form of annotation projection over parallel corpora or other multilingual resources or making use of transferable representations, such as phonetic transcriptions, closely related languages or bilingual dictionaries (Rahimi et al., 2019).

According to (Siddhant et al., 2020), there is still a sizable gap in the performance of cross-lingually transferred models, particularly on syntactic and sentence retrieval tasks when compared to the English Language. This makes a case for ground-up manually annotated and context-specific data for NLP tasks by native speakers. This work aims to explore training NER models for Nigerian English, Nigerian Pidgin English, Igbo, Hausa and Yoruba Languages.

2 Methodology

Figure 1 shows a simple flow for training Named Entity Recognition for the Nigerian Languages explored in this work.

![Methodology for training NER Models for Nigerian Languages](image)

2.1 Data Collection

50 articles were selected from a collection of online news articles for the 5 languages. The articles collected were in two major categories, General and Sports News with a split ratio of 4:1

2.2 Data Annotation

TagEditor software (v.2.3.2) an open-source annotation tool compatible with spaCy, was used by data annotators who are expert speakers of each language to label the named entities in their respective languages. These annotation gazettes were labelled based on the OntoNotes Entities and description as shown in Table 1. The annotated data for each language was then exported and formatted to the spaCy’s gold.doc_to_json. This is the acceptable data format for spaCy train.

| OntoNotes Named Entities | Description |
|--------------------------|-------------|
| PERSON                   | People, including fictional. |
| NORP                     | Nationalities or religious or political groups. |
| FAC                      | Buildings, airports, highways, bridges, etc. |
| ORG                      | Companies, agencies, institutions, etc. |
| GPE                      | Countries, cities, states, etc. |
| LOC                      | Names of geographical locations other than GPEs. |
| PRODUCT                  | Objects, vehicles, foods, etc. (Not services.) |
| EVENT                    | Named hurricanes, battles, wars, sports events, etc. |
| WORK, ART                | Titles of books, songs, etc. |
| LANGUAGE                 | Any named language. |
| DATE                     | Absolute or relative dates or periods. |
| TIME                     | Times smaller than a day. |
| PERCENT                  | Percentage, including %. |
| MONEY                    | Monetary values, including unit. |
| QUANTITY                 | Measurements, as of weight or distance. |
| ORDINAL                  | “first”, “second”, etc. |
| CARDINAL                 | Numerals that do not fall under another type. |

Table 1: spaCy OntoNotes Named Entities

2.3 NER Model Development

The approach for developing the models used in this research paper was based on spaCy. spaCy ¹ is an open-source software library for advanced natural language processing, written in the programming languages Python and Cython (Neumann et al., 2019).

spaCy’s Named Entity Recognition system features a sophisticated word embedding strategy using subword features and Bloom embedding strategy, a deep learning Convolutional Neural Network with residual connections and a novel transition-based approach (which involves layer normalization and maxout non-linearity) to Named Entity parsing, giving much better efficiency than the standard BiLSTM solution ².

spaCy takes in the annotated training dataset, which is the collection of documents labelled into existing entities. The algorithm uses machine learning to study patterns from the given set of training data.

For this project, the annotated data was split into 80% training and 20% evaluation sets. Table 2 represents number of sentences annotated (spaCy docs), which is different for every language. They were used while training the Named Entity Recognition(NER) models on the custom annotated data.

¹https://spacy.io/models
²https://www.youtube.com/watch?v=sqDHBH9IrU&t=2793s
for each of the languages. Igbo, Yoruba and Hausa NER models were trained on a blank spaCy model while the base NER models for Nigerian English and Pidgin ran on the spaCy “en” pretrained model. The latter was adopted since either of the language vocabularies had similarity with spaCy’s English NER Model. The metrics used in model assessment are Precision, Recall and F1-Score.

| NER Model       | No. of Training sentences | No. of Evaluation sentences |
|-----------------|---------------------------|-----------------------------|
| PidginNER       | 1148                      | 103                         |
| NigerianEnglishNER | 512                     | 95                          |
| IgboNER         | 502                       | 196                         |
| YorubaNER       | 549                       | 197                         |
| HausaNER        | 2658                      | 1046                        |
| NaajaNER        | 5697                      | 1135                        |

Table 2: NER Model Training Parameters

3 Results

Tables 3—9 show the detailed number of entities, entity classification F1 score, Precision and Recall scores for each entity for individual language models and the combined NER model. Each table is named with the language then a suffix—NER.

Results for 50 annotated articles (80% General News, 20% Sports News)

| NER Model       | Total Number of words | F1 Score | Precision | Recall |
|-----------------|-----------------------|----------|-----------|--------|
| PidginNER       | 24716                 | 67.13    | 70.29     | 64.24  |
| NigerianEnglishNER | 7246               | 54.49    | 55.48     | 53.53  |
| IgboNER         | 18577                 | 35.48    | 47.18     | 28.43  |
| YorubaNER       | 17169                 | 45.57    | 53.64     | 39.61  |
| HausaNER        | 20265                 | 44.84    | 52.59     | 39.08  |
| NaajaNER        | 87710                 | 53.75    | 62.43     | 47.19  |

Table 3: Precision, Recall and F1 Scores for each model

4 Discussion

Models developed are available here\(^3\) with an access to an interactive web app available here\(^4\).

Tables 4–8 shows the metric scores observed for each of the NER models on the test data. There are certain entities in the entire dataset with low representation (less than 50), which resulted to the lower performance of the models when identifying such entities in the given test data. Therefore >50 was used for better performance.

The PidginNER and NigerianEnglishNER models were trained by leveraging the pretrained spaCy

\(^3\)https://git.io/JYo0kk

\(^4\)https://nigner.herokuapp.com/
Table 7: Entity Evaluation Scores of the PidginNER Model

| Entity Type | No. of examples per entity | F1 Score | Precision | Recall |
|-------------|---------------------------|----------|-----------|--------|
| CARDINAL    | >50                       | 82.60    | 80.85     | 84.44  |
| DATE        | >50                       | 68.49    | 65.79     | 71.43  |
| EVENT       | >50                       | 71.43    | 71.43     | 71.43  |
| FAC         | >50                       | 0.00     | 0.00      | 0.00   |
| GPE         | >50                       | 53.33    | 80.00     | 40.00  |
| LANGUAGE    | 5                         | 0.00     | 0.00      | 0.00   |
| LAW         | 4                         | 0.00     | 0.00      | 0.00   |
| LOC         | 5                         | 0.00     | 0.00      | 0.00   |
| MONEY       | >50                       | 0.00     | 0.00      | 0.00   |
| NORP        | >50                       | 22.22    | 12.5      | 100.00 |
| ORDINAL     | >50                       | 100.00   | 100.00    | 100.00 |
| ORG         | >50                       | 65.15    | 70.00     | 60.87  |
| PERCENT     | 8                         | 0.00     | 0.00      | 0.00   |
| PERSON      | >50                       | 67.94    | 71.72     | 64.55  |
| PRODUCT     | 14                        | 0.00     | 0.00      | 0.00   |
| QUANTITY    | 45                        | 50.00    | 33.33     | 100.00 |
| TIME        | 21                        | 66.67    | 100.00    | 50.00  |
| WORK_OF_ART | >50                       | 0.00     | 0.00      | 0.00   |

Table 8: Entity Evaluation Scores of the NigerianEnglishNER Model

| Entity Type | No. of examples per entity | F1 Score | Precision | Recall |
|-------------|---------------------------|----------|-----------|--------|
| CARDINAL    | >50                       | 75.00    | 70.77     | 70.00  |
| DATE        | >50                       | 59.26    | 57.14     | 61.54  |
| EVENT       | >50                       | 42.11    | 57.14     | 33.33  |
| FAC         | 7                         | 0.00     | 0.00      | 0.00   |
| GPE         | >50                       | 69.84    | 68.75     | 70.97  |
| LANGUAGE    | 2                         | 0.00     | 0.00      | 0.00   |
| LAW         | 5                         | 0.00     | 0.00      | 0.00   |
| MONEY       | 5                         | 0.00     | 0.00      | 0.00   |
| NORP        | >50                       | 63.64    | 53.85     | 77.78  |
| ORDINAL     | 13                        | 66.67    | 66.67     | 66.67  |
| ORG         | >50                       | 29.10    | 25.81     | 33.33  |
| PERCENT     | 8                         | 0.00     | 0.00      | 0.00   |
| PERSON      | >50                       | 43.14    | 50.00     | 37.93  |
| PRODUCT     | 14                        | 0.00     | 0.00      | 0.00   |
| QUANTITY    | >50                       | 50.00    | 33.33     | 100.00 |
| TIME        | 4                         | 0.00     | 0.00      | 0.00   |
| WORK_OF_ART | 3                         | 0.00     | 0.00      | 0.00   |

Table 9: Entity Evaluation Scores of the NaijaNER Model

| Entity Type | No. of examples per entity | F1 Score | Precision | Recall |
|-------------|---------------------------|----------|-----------|--------|
| CARDINAL    | >50                       | 61.66    | 57.35     | 66.67  |
| DATE        | >50                       | 63.47    | 57.14     | 61.54  |
| EVENT       | >50                       | 38.41    | 65.91     | 27.10  |
| FAC         | >50                       | 0.00     | 0.00      | 0.00   |
| GPE         | >50                       | 60.10    | 68.75     | 54.84  |
| LANGUAGE    | >50                       | 85.71    | 85.71     | 85.71  |
| LAW         | 28                        | 0.00     | 0.00      | 0.00   |
| LOC         | >50                       | 8.70     | 25.00     | 5.26   |
| MONEY       | >50                       | 0.00     | 0.00      | 0.00   |
| NORP        | >50                       | 25.64    | 50.00     | 17.24  |
| ORDINAL     | >50                       | 36.36    | 53.85     | 27.45  |
| ORG         | >50                       | 64.63    | 61.86     | 67.67  |
| PERCENT     | 28                        | 0.00     | 0.00      | 0.00   |
| PERSON      | >50                       | 59.88    | 66.60     | 54.38  |
| PRODUCT     | >50                       | 4.68     | 20.00     | 2.65   |
| QUANTITY    | >50                       | 38.46    | 35.08     | 42.55  |
| TIME        | >50                       | 58.82    | 83.33     | 45.45  |
| WORK_OF_ART | >50                       | 20.57    | 30.91     | 13.74  |

In the Igbo sample, the NaijaNER model recognized Ukochukwu Elliot Ugochukwu Uko as PER-
SON, which was a better performance to the IgboNER model. Also, for the Yoruba sample text, and just like the YorubaNER model 6, the NaijaNER model 7 had some misidentifications of entities such as "Lori" (on), "Pdp" (a political party) and "Daily Sun" (a News company) as PERSON, whereas the rest of the entities were correctly named. Lastly, for the Hausa language, the NaijaNER model 9 was accurate in comparison to the HausaNER model 8, except for the unnecessary classification of "dan jaridar" (journalist) as ORG and GPE.

**Figure 7: NaijaNER on Yoruba Text**

**Figure 8: HausaNER on Hausa Text**

**Figure 9: NaijaNER on Hausa Text**

**Figure 10: NigeriaEnglishNER on English text in Nigerian context**

**Figure 11: spaCy English NER on English text in Nigerian context**

These models are relevant to the context that Nigerian sentences have. This can be attributed to the purpose of developing the NigerianEnglishNER model. Figure 10 shows that the NigerianEnglishNER model performed well in correctly identifying the entities in the text including "Imo State" (a State in Nigeria), that the spaCy en model 11 could not recognise.

## 5 Conclusion

This work explores individual and combined Named Entity Recognition in 5 Nigerian languages - Nigerian English, Pidgin, Yoruba, Hausa and Igbo. The choice of when to use a combined or individual models is dependent on the user’s interest. The combined NER model-NaijaNER, is versatile for multi-language applications and has a good performance on each of the languages while individual language models showed better performance on their specific languages but will require deployment of multiple models.

Combined Named Entity Recognition that works with different languages can be a valuable tool in low resourced language systems and can help drive inclusion, ease of deployment in production and reusability of NER models.

## References

Bing Li. 2021. Named entity recognition in the style of object detection. *arXiv preprint arXiv:2101.11122*.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named entity recognition without gazetteers.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguistics Investigations*, 30(1):3–26.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. *arXiv preprint arXiv:1902.00193*.

Alan Ritter, Sam Clark, Mauzam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. pages 1524–1534.

Sciforce. 2019. Nlp for low-resource settings.

Yusuke Shinyama and Satoshi Sekine. 2004. Named entity discovery using comparable news articles. pages 848–853.

Aditya Siddhant, Junjie Hu, Melvin Johnson, Orhan Firat, and Sebastian Ruder. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.