MasakhaNER: Named Entity Recognition for African Languages

David Ifeoluwa Adelani[1*], Jade Abbott[2*], Graham Neubig[3], Daniel D'souza[4*], Julia Kreutzer[5*], Constantine Lignos[6*], Chester Palen-Michel[6*], Happy Buzaaba[7*], Shruti Rijhwani[3], Sebastian Ruder[8], Stephen Mayhew[9], Israel Abebe Azime[10*], Shamsuddeen H. Muhammad[11,12*], Chris Chinenye Emezue[13*], Joyce Nakatumba-Nabende[14*], Perez Ogayo[15*], Aremu Anuoluwapo[16*], Catherine Gitau[*], Derguene Mbaye[*], Jesujoba Alabi[17*], Seid Muhie Yimam[18], Tajuddeen Rabiu Gwadabe[19*], Ignatius Ezeani[20*], Rubungo Andre Niyongabo[21*], Jonathan Mukiibi[14], Verrah Otiende[22*], Iroro Orife[23*], Davis David[*], Samba Ngom[*], Tosin Adewumi[24*], Paul Rayson[20], Mofetoluwa Adeyemi[*], Gerald Muriuki[14], Emmanuel Anebi[*], Chiamaka Chukwuneke[20], Nkiruka Odu[25], Eric Peter Wairagala[14], Samuel Oyerinde[*], Clemencia Siro[*], Tobius Saul Bateesa[14], Temilola Oloyede[*], Yvonne Wambui[*], Victor Akinode[*], Deborah Nabagereka[14], Maurice Katusiime[14], Ayodele Awokoya[26*], Mouhamadane MBOUP[*], Dibora Gebreyohannes[*], Henok Tilaye[*], Kelechi Nwaike[*], Degaga Wolde[*], Abdoulaye Faye[*], Blessing Sibanda[27*], Orevaoghene Ahia[28*], Bonaventure F. P. Dossou[29*], Kelechi Ogueji[30*], Thierno Ibrahima DIOP[*], Abdoulaye Diallo[*], Adewale Akinfaderin[*], Tendai Marengereke[*], and Salomey Osei[10*]

[*]Masakhane NLP, [1]Spoken Language Systems Group (LSV), Saarland University, Germany, [2]Retro Rabbit, South Africa, [3]Language Technologies Institute, Carnegie Mellon University, United States, [4]ProQuest, United States, [5]Google Research, Canada, [6]Brandeis University, United States, [8]DeepMind, United Kingdom, [9]Duolingo, United States, [7]Graduate School of Systems and Information Engineering, University of Tsukuba, Japan, [10]African Institute for Mathematical Sciences (AIMS-AMMI), Ethiopia, [11]University of Porto, Nigeria, [12]Bayero University, Kano, Nigeria, [13]Technical University of Munich, Germany [14] Makerere University, Kampala, Uganda,[15]African Leadership University, Rwanda [16]University of Lagos, Nigeria, [17]Max Planck Institute for Informatics, Germany, [18]LT Group, Universität Hamburg, Germany, [19]University of Chinese Academy of Science, China [20]Lancaster University, United Kingdom, [21]University of Electronic Science and Technology of China, China, [22]United States International University - Africa (USIU-A), Kenya, [23]Niger-Volta LTI [24]Luleå University of Technology, Sweden [25]African University of Science and Technology, Abuja, Nigeria [26]University of Ibadan, Nigeria, [27]Namibia University of Science and Technology, Namibia [28]Instadeep, Nigeria [29]Jacobs University Bremen, Germany, [30] University of Waterloo, Canada

Abstract

We take a step towards addressing the under-representation of the African continent in NLP research by bringing together different stakeholders to create the first large, publicly available, high-quality dataset for named entity recognition (NER) in ten African languages. We detail the characteristics of these languages to help researchers and practitioners better understand the challenges they pose for NER tasks. We analyze our datasets and conduct an extensive empirical evaluation of state-of-the-art methods across both supervised and transfer learning settings. Finally, we release the data, code, and models to inspire future research on African NLP.[1]

[1]https://git.io/masakhane-ner.

1 Introduction

Africa has over 2,000 spoken languages (Eberhard et al., 2020); however, these languages are scarcely represented in existing natural language processing (NLP) datasets, research, and tools (Martinus and Abbott, 2019). ∀ (2020) investigate the reasons for these disparities by examining how NLP for low-resource languages is constrained by several societal factors. One of these factors is the geographical and language diversity of NLP researchers. For example, of the 2695 affiliations of authors whose works were published at the five major NLP conferences in 2019, only five were from African institutions (Caines, 2019). Conversely, many NLP tasks such

Transactions of the Association for Computational Linguistics, vol. 9, pp. 1116–1131, 2021. https://doi.org/10.1162/tacl_a_00416
Action Editor: Miguel Ballesteros. Submission batch: 5/2021; Revision batch: 7/2021; Published 10/2021.

as machine translation, text classification, part-of-speech tagging, and named entity recognition would benefit from the knowledge of native speakers who are involved in the development of datasets and models.

In this work, we focus on named entity recognition (NER)—one of the most impactful tasks in NLP (Sang and De Meulder, 2003; Lample et al., 2016). NER is an important information extraction task and an essential component of numerous products including spell-checkers, localization of voice and dialogue systems, and conversational agents. It also enables identifying African names, places, and organizations for information retrieval. African languages are underrepresented in this crucial task due to lack of datasets, reproducible results, and researchers who understand the challenges that such languages present for NER.

In this paper, we take an initial step towards improving representation for African languages for the NER task, making the following contributions:

(i) We bring together language speakers, dataset curators, NLP practitioners, and evaluation experts to address the challenges facing NER for African languages. Based on the availability of online news corpora and language annotators, we develop NER datasets, models, and evaluation covering ten widely spoken African languages.

(ii) We curate NER datasets from local sources to ensure relevance of future research for native speakers of the respective languages.

(iii) We train and evaluate multiple NER models for all ten languages. Our experiments provide insights into the transfer across languages, and highlight open challenges.

(iv) We release the datasets, code, and models to facilitate future research on the specific challenges raised by NER for African languages.

2 Related Work

African NER Datasets NER is a well-studied sequence labeling task (Yadav and Bethard, 2018) and has been the subject of many shared tasks in different languages (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003; Sangal et al., 2008; Shaalan, 2014; Benikova et al., 2014). However, most of the available datasets are in high-resource languages. Although there have been efforts to create NER datasets for lower-resourced languages, such as the WikiAnn corpus (Pan et al., 2017) covering 282 languages, such datasets consist of ''silver-standard'' labels created by transferring annotations from English to other languages through cross-lingual links in knowledge bases. Because the WikiAnn corpus data comes from Wikipedia, it includes some African languages; though most have fewer than 10k tokens.

Other NER datasets for African languages include SADiLaR (Eiselen, 2016) for ten South African languages based on government data, and small corpora of fewer than 2K sentences for Yorùbá (Alabi et al., 2020) and Hausa (Hedderich et al., 2020). Additionally, the LORELEI language packs (Strassel and Tracey, 2016) include some African languages (Yorùbá, Hausa, Amharic, Somali, Twi, Swahili, Wolof, Kinyarwanda, and Zulu), but are not publicly available.

NER Models Popular sequence labeling models for NER include the CRF (Lafferty et al., 2001), CNN-BiLSTM (Chiu and Nichols, 2016), BiLSTM-CRF (Huang et al., 2015), and CNN-BiLSTM-CRF (Ma and Hovy, 2016). The traditional CRF makes use of hand-crafted features like part-of-speech tags, context words and word capitalization. Neural NER models on the other hand are initialized with word embeddings like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017). More recently, pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and LUKE (Yamada et al., 2020) have been applied to produce state-of-the-art results for the NER task. Multilingual variants of these models like mBERT and XLM-RoBERTa (Conneau et al., 2020) make it possible to train NER models for several languages using transfer learning. Language-specific parameters and adaptation to unlabeled data of the target language have yielded further gains (Pfeiffer et al., 2020a,b).

3 Focus Languages

Table 1 provides an overview of the languages considered in this work, their language family,

Language	Family	Speakers	Region
Amharic	Afro-Asiatic-Ethio-Semitic	33M	East
Hausa	Afro-Asiatic-Chadic	63M	West
Igbo	Niger-Congo-Volta-Niger	27M	West
Kinyarwanda	Niger-Congo-Bantu	12M	East
Luganda	Niger-Congo-Bantu	7M	East
Luo	Nilo Saharan	4M	East
Nigerian-Pidgin	English Creole	75M	West
Swahili	Niger-Congo-Bantu	98M	Central & East
Wolof	Niger-Congo-Senegambia	5M	West & NW
Yorùbá	Niger-Congo-Volta-Niger	42M	West

Table 1: Language, family, number of speakers (Eberhard et al., 2020), and regions in Africa.

number of speakers and the regions in Africa where they are spoken. We chose to focus on these languages due to the availability of online news corpora, annotators, and most importantly because they are widely spoken native African languages. Both region and language family might indicate a notion of proximity for NER, either because of linguistic features shared within that family, or because data sources cover a common set of locally relevant entities. We highlight language specifics for each language to illustrate the diversity of this selection of languages in Section 3.1, and then showcase the differences in named entities across these languages in Section 3.2.

3.1 Language Characteristics

Amharic (amh) uses the Fidel script consisting of 33 basic scripts (ሀ (hä) ለ (lä) መ (mä) ሠ (šä)...), each of them with at least 7 vowel sequences (such as ሀ (hä) ሁ (hu) ሂ (hi) ሃ (ha) ሄ (he) ህ (hi) ሆ (ho)). This results in more than 231 characters or Fidels. Numbers and punctuation marks are also represented uniquely with specific Fidels (፩ **(1)**, ፪ **(2)**, ... and ። **(.),** ፨ **(!),** ፤ **(;),**).

Hausa (hau) has 23–25 consonants, depending on the dialect and five short and five long vowels. Hausa has labialized phonemic consonants, as in /gw/ (e.g., 'agwagwa'). As found in some African languages, implosive consonants also exist in Hausa (e.g., 'b, 'd, etc., as in 'barna'). Similarly, the Hausa approximant 'r' is realized in two distinct manners: roll and trill, as in 'rai' and 'ra'ayi', respectively.

Igbo (ibo) is an agglutinative language, with many frequent suffixes and prefixes (Emenanjo, 1978). A single stem can yield many word-forms by addition of affixes that extend its original meaning (Onyenwe and Hepple, 2016). Igbo is also tonal, with two distinctive tones (high and low) and a down-stepped high tone in some cases. The alphabet consists of 28 consonants and 8 vowels (A, E, I, Ị, O, Ọ, U, Ụ). In addition to the Latin letters (except *c*), Igbo contains the following digraphs: (ch, gb, gh, gw, kp, kw, nw, ny, sh).

Kinyarwanda (kin) makes use of 24 Latin characters with 5 vowels similar to English and 19 consonants excluding q and x. Moreover, Kinyarwanda has 74 additional complex consonants (such as mb, mpw, and njyw) (Government, 2014). It is a tonal language with three tones: low (no diacritic), high (signaled by ''/''), and falling (signaled by ''^''). The default word order is subject-verb-object.

Luganda (lug) is a tonal language with subject-verb-object word order. The Luganda alphabet is composed of 24 letters that include 17 consonants (p, v, f, m, d, t, l, r, n, z, s, j, c, g), 5 vowel sounds represented in the five alphabetical symbols (a, e, i, o, u), and 2 semi-vowels (w, y). It also has a special consonant ng'.

Luo (luo) is a tonal language with 4 tones (high, low, falling, rising), although the tonality is not marked in orthography. It has 26 Latin consonants without Latin letters (c, q, v, x, and z) and additional consonants (ch, dh, mb, nd, ng', ng, ny, nj, th, sh). There are nine vowels (a, e, i, o, u, ɐ, ɛ, ɔ, ʊ) which are distinguished primarily by advanced tongue root (ATR) harmony (De Pauw et al., 2007).

Nigerian-Pidgin (pcm) is a largely oral, national lingua franca with a distinct phonology from English, its lexifier language. Portuguese, French, and especially indigenous languages form the substrate of lexical, phonological, syntactic, and semantic influence on Nigerian-Pidgin (NP). English lexical items absorbed by NP are often phonologically closer to indigenous Nigerian languages, notably in the realization of vowels. As a rapidly evolving language, the NP orthography is undergoing codification and indigenization (Offiong Mensah, 2012; Onovbiona, 2012; Ojarikre, 2013).

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጄርያ ፩፰ ዓመት ያሳለፈውን ዛንግን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Najeriya sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke nọgoro afọ iri na asatọ na Naijiria
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Ẹ́míà ìlú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjidínlógún ní orílè-èdè Nàìjíríà

Table 2: Example of named entities in different languages. PER, LOC, and DATE are in colours purple, orange, and green, respectively.

Swahili (swa) is the most widely spoken language on the African continent. It has 30 letters including 24 Latin letters without characters (q and x) and six additional consonants (ch, dh, gh, ng', sh, th) unique to Swahili pronunciation.

Wolof (wol) has an alphabet similar to that of French. It consists of 29 characters, including all letters of the French alphabet except h, v, and z. It also includes the characters Ŋ (''ng'', lowercase: ŋ) and Ñ' (''gn'' as in Spanish). Accents are present, but limited in number (À, É, Ë, Ó). However, unlike many other Niger-Congo languages, Wolof is not a tonal language.

Yorùbá (yor) has 25 Latin letters without the Latin characters (c, q, v, x, and z) and with additional letters (ẹ, gb, ṣ, ọ). Yorùbá is a tonal language with three tones: low (''\''), middle (''–'', optional) and high (''/''). The tonal marks and underdots are referred to as diacritics and they are needed for the correct pronunciation of a word. Yorùbá is a highly isolating language and the sentence structure follows subject-verb-object.

3.2 Named Entities

Most of the work on NER is centered around English, and it is unclear how well existing models can generalize to other languages in terms of sentence structure or surface forms. In Hu et al.'s (2020) evaluation on cross-lingual generalization for NER, only two African languages were considered and it was seen that transformer-based models particularly struggled to generalize to named entities in Swahili. To highlight the differences across our focus languages, Table 2 shows an English[2] example sentence, with color-coded PER, LOC, and DATE entities, and the corresponding translations. The following characteristics of the languages in our dataset could pose challenges for NER systems developed for English:

- Amharic shares no lexical overlap with the English source sentence.
- While ''Zhang'' is identical across all Latin-script languages, ''Kano'' features accents in Wolof and Yorùbá due to its localization.
- The Fidel script has no capitalization, which could hinder transfer from other languages.
- Igbo, Wolof, and Yorùbá all use diacritics, which are not present in the English alphabet.
- The surface form of named entities (NE) is the same in English and Nigerian-Pidgin, but there exist lexical differences (e.g., in terms of how time is realized).
- Between the 10 African languages,''Nigeria'' is spelled in 6 different ways.
- Numerical ''18'': Igbo, Wolof, and Yorùbá write out their numbers, resulting in different numbers of tokens for the entity span.

[2]Although the original sentence is from BBC Pidgin https://www.bbc.com/pidgin/tori-51702073.

Language	Data Source	Train/ dev/ test	# Anno.	PER	ORG	LOC	DATE	% of Entities in Tokens	# Tokens
Amharic	DW & BBC	1750/ 250/ 500	4	730	403	1,420	580	15.13	37,032
Hausa	VOA Hausa	1903/ 272/ 545	3	1,490	766	2,779	922	12.17	80,152
Igbo	BBC Igbo	2233/ 319/ 638	6	1,603	1,292	1,677	690	13.15	61,668
Kinyarwanda	IGIHE news	2110/ 301/ 604	2	1,366	1,038	2,096	792	12.85	68,819
Luganda	BUKEDDE news	2003/ 200/ 401	3	1,868	838	943	574	14.81	46,615
Luo	Ramogi FM news	644/ 92/ 185	2	557	286	666	343	14.95	26,303
Nigerian-Pidgin	BBC Pidgin	2100/ 300/ 600	5	2,602	1,042	1,317	1,242	13.25	76,063
Swahili	VOA Swahili	2104/ 300/ 602	6	1,702	960	2,842	940	12.48	79,272
Wolof	Lu Defu Waxu & Saabal	1,871/ 267/ 536	2	731	245	836	206	6.02	52,872
Yorùbá	GV & VON news	2124/ 303/ 608	5	1,039	835	1,627	853	11.57	83,285

Table 3: Statistics of our datasets including their source, number of sentences in each split, number of annotators, number of entities of each label type, percentage of tokens that are named entities, and total number of tokens.

4 Data and Annotation Methodology

Our data were obtained from local news sources, in order to ensure relevance of the dataset for native speakers from those regions. The dataset was annotated using the ELISA tool (Lin et al., 2018) by native speakers who come from the same regions as the news sources and volunteered through the *Masakhane* community.[3] Annotators were not paid but are all included as authors of this paper. The annotators were trained on how to perform NER annotation using the MUC-6 annotation guide.[4] We annotated four entity types: Personal name (`PER`), Location (`LOC`), Organization (`ORG`), and date & time (`DATE`). The annotated entities were inspired by the English CoNLL-2003 Corpus (Tjong Kim Sang, 2002). We replaced the `MISC` tag with the `DATE` tag following Alabi et al. (2020) as the `MISC` tag may be ill-defined and cause disagreement among non-expert annotators. We report the number of annotators as well as general statistics of the datasets in Table 3. For each language, we divided the annotated data into training, development, and test splits consisting of 70%, 10%, and 20% of the data, respectively.

A key objective of our annotation procedure was to create high-quality datasets by ensuring high annotator agreement. To achieve high agreement scores, we ran collaborative workshops for each language, which allowed annotators to discuss any disagreements. ELISA provides an entity-level F1-score and also an interface for annotators to correct their mistakes, making it easy to achieve inter-annotator agreement scores between 0.96 and 1.0 for all languages.

Dataset	Token Fleiss' κ	Entity Fleiss' κ	Disagreement from Type
amh	0.987	0.959	0.044
hau	0.988	0.962	0.097
ibo	0.995	0.983	0.071
kin	1.000	1.000	0.000
lug	0.997	0.990	0.023
luo	1.000	1.000	0.000
pcm	0.989	0.966	0.048
swa	1.000	1.000	0.000
wol	1.000	1.000	0.000
yor	0.990	0.964	0.079

Table 4: Inter-annotator agreement for our datasets calculated using Fleiss' kappa (κ) at the token and entity level. Disagreement from type refers to the proportion of all entity-level disagreements, which are due only to type mismatch.

We report inter-annotator agreement scores in Table 4 using Fleiss' kappa (Fleiss, 1971) at both the token and entity level. The latter considers each span an annotator proposed as an entity. As a result of our workshops, all our datasets have exceptionally high inter-annotator agreement. For Kinyarwanda, Luo, Swahili, and Wolof, we report perfect inter-annotator agreement scores ($\kappa = 1$). For each of these languages, two annotators annotated each token and were instructed to discuss and resolve conflicts among themselves. The Appendix provides a detailed entity-level confusion matrix in Table 11.

[3] `https://www.masakhane.io`.

[4] `https://cs.nyu.edu/~grishman/muc6.html`.

5 Experimental Setup

5.1 NER Baseline Models

To evaluate baseline performance on our dataset, we experiment with three popular NER models: CNN-BiLSTM-CRF, multilingual BERT (mBERT), and XLM-RoBERTa (XLM-R). The latter two models are implemented using the HuggingFace transformers toolkit (Wolf et al., 2019). For each language, we train the models on the in-language training data and evaluate on its test data.

CNN-BiLSTM-CRF This architecture was proposed for NER by Ma and Hovy (2016). For each input sequence, we first compute the vector representation for each word by concatenating character-level encodings from a CNN and vector embeddings for each word. Following Rijhwani et al. (2020), we use randomly initialized word embeddings since we do not have high-quality pre-trained embeddings for all the languages in our dataset. Our model is implemented using the DyNet toolkit (Neubig et al., 2017).

mBERT We fine-tune multilingual BERT (Devlin et al., 2019) on our NER corpus by adding a linear classification layer to the pre-trained transformer model, and train it end-to-end. mBERT was trained on 104 languages including only two African languages: Swahili and Yorùbá. We use the mBERT-base cased model with 12-layer Transformer blocks consisting of 768-hidden size and 110M parameters.

XLM-R XLM-R (Conneau et al., 2020) was trained on 100 languages including Amharic, Hausa, and Swahili. The major differences between XLM-R and mBERT are (1) XLM-R was trained on Common Crawl while mBERT was trained on Wikipedia; (2) XLM-R is based on RoBERTa, which is trained with a masked language model (MLM) objective while mBERT was additionally trained with a next sentence prediction objective. We make use of the XLM-R base and large models for the baseline models. The XLM-R-base model consisting of 12 layers, with a hidden size of 768 and 270M parameters. On the other hand, the XLM-R-large has 24 layers, with a hidden size of 1024 and 550M parameters.

MeanE-BiLSTM This is a simple BiLSTM model with an additional linear classifier. For each input sequence, we first extract a sentence embedding from mBERT or XLM-R language model (LM) before passing it into the BiLSTM model. Following Reimers and Gurevych (2019), we make use of the mean of the 12-layer output embeddings of the LM (i.e., *MeanE*). This has been shown to provide better sentence representations than the embedding of the `[CLS]` token used for fine-tuning mBERT and XLM-R.

Language BERT The mBERT and the XLM-R models only support two and three languages under study, respectively. One effective approach to adapt the pre-trained transformer models to new domains is ‘‘domain-adaptive fine-tuning’’ (Howard and Ruder, 2018; Gururangan et al., 2020)—fine-tuning on unlabeled data in the new domain, which also works very well when adapting to a new language (Pfeiffer et al., 2020a; Alabi et al., 2020). For each of the African languages, we performed *language-adaptive fine-tuning* on available unlabeled corpora mostly from JW300 (Agić and Vulić, 2019), indigenous news sources, and XLM-R Common Crawl corpora (Conneau et al., 2020). The Appendix provides the details of the unlabeled corpora in Table 10. This approach is quite useful for languages whose scripts are not supported by the multi-lingual transformer models like Amharic where we replace the vocabulary of mBERT by an Amharic vocabulary before we perform language-adaptive fine-tuning, similar to Alabi et al. (2020).

5.2 Improving the Baseline Models

In this section, we consider techniques to improve the baseline models such as utilizing gazetteers, transfer learning from other domains, and languages, and aggregating NER datasets by regions. For these experiments, we focus on the `PER`, `ORG`, and `LOC` categories, because the gazetteers from Wikipedia do not contain `DATE` entities and some source domains and languages that we transfer from do not have the `DATE` annotation. We apply these modifications to the XLM-R model because it generally outperforms mBERT in our experiments (see Section 6).

5.2.1 Gazetteers for NER

Gazetteers are lists of named entities collected from manually crafted resources such as

GeoNames or Wikipedia. Before the widespread adoption of neural networks, NER methods used gazetteers-based features to improve performance (Ratinov and Roth, 2009). These features are created for each n-gram in the dataset and are typically binary-valued, indicating whether that n-gram is present in the gazetteer.

Recently, Rijhwani et al. (2020) showed that augmenting the neural CNN-BiLSTM-CRF model with gazetteer features can improve NER performance for low-resource languages. We conduct similar experiments on the languages in our dataset, using entity lists from Wikipedia as gazetteers. For Luo and Nigerian-Pidgin, which do not have their own Wikipedia, we use entity lists from English Wikipedia.

5.2.2 Transfer Learning

Here, we focus on cross-domain transfer from Wikipedia to the news domain, and cross-lingual transfer from English and Swahili NER datasets to the other languages in our dataset.

Domain Adaptation from WikiAnn We make use of the WikiAnn corpus (Pan et al., 2017), which is available for five of the languages in our dataset: Amharic, Igbo, Kinyarwanda, Swahili, and Yorùbá. For each language, the corpus contains 100 sentences in each of the training, development and test splits except for Swahili, which contains 1K sentences in each split. For each language, we train on the corresponding WikiAnn training set and either zero-shot transfer to our respective test set or additionally fine-tune on our training data.

Cross-lingual Transfer For training the cross-lingual transfer models, we use the CoNLL-2003[5] NER dataset in English with over 14K training sentences and our annotated corpus. The reason for CoNLL-2003 is because it is in the same news domain as our annotated corpus. We also make use of the languages that are supported by the XLM-R model and are widely spoken in East and West Africa like Swahili and Hausa. The English corpus has been shown to transfer very well to low-resource languages (Hedderich et al., 2020; Lauscher et al., 2020). We first train on either the English CoNLL-2003 data or our training data in Swahili, Hausa, or Nigerian-Pidgin before testing on the target African languages.

[5]We also tried OntoNotes 5.0 by combining FAC & ORG as ''ORG'' and GPE & LOC as ''LOC'' and others as ''O'' except ''PER'', but it gave lower performance in zero-shot transfer (19.38 F1) while CoNLL-2003 gave 37.15 F1.

5.3 Aggregating Languages by Regions

As previously illustrated in Table 2, several entities have the same form in different languages while some entities may be more common in the region where the language is spoken. To study the performance of NER models across geographical areas, we combine languages based on the region of Africa that they are spoken in (see Table 1): (1) East region with Kinyarwanda, Luganda, Luo, and Swahili; (2) West Region with Hausa, Igbo, Nigerian-Pidgin, Wolof, and Yorùbá languages, (3) East and West regions—all languages except Amharic because of its distinct writing system.

6 Results

6.1 Baseline Models

Table 5 gives the F1-score obtained by CNN-BiLSTM-CRF, mBERT, and XLM-R models on the test sets of the ten African languages when training on our in-language data. We additionally indicate whether the language is supported by the pre-trained language models (✓). The percentage of entities that are of out-of-vocabulary (OOV; entities in the test set that are not present in the training set) is also reported alongside results of the baseline models. In general, the datasets with greater numbers of OOV entities have lower performance with the CNN-BiLSTM-CRF model, while those with lower OOV rates (Hausa, Igbo, Swahili) have higher performance. We find that the CNN-BiLSTM-CRF model performs worse than fine-tuning mBERT and XLM-R models end-to-end (FTune). We expect performance to be better (e.g., for Amharic and Nigerian-Pidgin with over 18 F1 point difference) when using pre-trained word embeddings for the initialization of the BiLSTM model rather than random initialization (we leave this for future work as discussed in Section 7).

Interestingly, the pre-trained language models (PLMs) have reasonable performance even on languages they were not trained on such as Igbo, Kinyarwanda, Luganda, Luo, and Wolof. However, languages supported by the PLM tend

Lang.	In mBERT?	In XLM-R?	% OOV in Test Entities	CNN-BiLSTM CRF	mBERT-base MeanE / FTune	XLM-R-base MeanE / FTune	XLM-R Large FTune	lang. BERT FTune	lang. XLM-R FTune
amh	✗	✓	72.94	52.08	0.0 / 0.0	63.57 / 70.62	76.18	60.89	**77.97**
hau	✗	✓	33.40	83.52	81.49 / 86.65	86.06 / 89.50	90.54	91.31	**91.47**
ibo	✗	✗	46.56	80.02	76.17 / 85.19	73.47 / 84.78	84.12	86.75	**87.74**
kin	✗	✗	57.85	62.97	65.85 / 72.20	63.66 / 73.32	73.75	77.57	**77.76**
lug	✗	✗	61.12	74.67	70.38 / 80.36	68.15 / 79.69	81.57	83.44	**84.70**
luo	✗	✗	65.18	65.98	56.56 / 74.22	52.57 / 74.86	73.58	**75.59**	75.27
pcm	✗	✗	61.26	67.67	81.87 / 87.23	81.93 / 87.26	89.02	89.95	**90.00**
swa	✓	✓	40.97	78.24	83.08 / 86.80	84.33 / 87.37	89.36	89.36	**89.46**
wol	✗	✗	69.73	59.70	57.21 / 64.52	54.97 / 63.86	67.90	**69.43**	68.31
yor	✓	✗	65.99	67.44	74.28 / 78.97	67.45 / 78.26	78.89	82.58	**83.66**
avg	–	–	57.50	69.23	64.69 / 71.61	69.62 / 78.96	80.49	80.69	**82.63**
avg (excl. amh)	–	–	55.78	71.13	71.87 / 79.88	70.29 / 79.88	80.97	82.89	**83.15**

Table 5: NER model comparison, showing F1-score on the test sets after 50 epochs averaged over 5 runs. This result is for all 4 tags in the dataset: PER, ORG, LOC, DATE. **Bold** marks the top score (tied if within the range of SE). mBERT and XLM-R are trained in two ways (1) MeanE: mean output embeddings of the 12 LM layers are used to initialize BiLSTM + Linear classifier, and (2) FTune: LM fine-tuned end-to-end with a linear classifier. Lang. BERT & Lang XLM-R (base) are models fine-tuned after language adaptive fine-tuning.

to have better performance overall. We observe that fine-tuned XLM-R-base models have significantly better performance on five languages; two of the languages (Amharic and Swahili) are supported by the pre-trained XLM-R. Similarly, fine-tuning mBERT has better performance for Yorùbá since the language is part of the PLM's training corpus. Although mBERT is trained on Swahili, XLM-R-base shows better performance. This observation is consistent with Hu et al. (2020) and could be because XLM-R is trained on more Swahili text (Common Crawl with 275M tokens) whereas mBERT is trained on a smaller corpus from Wikipedia (6M tokens[6]).

Another observation is that mBERT tends to have better performance for the non-Bantu Niger-Congo languages (i.e., Igbo, Wolof, and Yorùbá) while XLM-R-base works better for Afro-Asiatic languages (i.e., Amharic and Hausa), Nilo-Saharan (i.e., Luo), and Bantu languages like Kinyarwanda and Swahili. We also note that the writing script is one of the primary factors influencing the transfer of knowledge in PLMs with regard to the languages they were not trained on. For example, mBERT achieves an F1-score of 0.0 on Amharic because it has not encountered the script during pre-training. In general, we find the fine-tuned XLM-R-large (with 550M parameters) to be better than XLM-R-base (with 270M parameters) and mBERT (with 110 parameters) in almost all languages. However, mBERT models perform slightly better for Igbo, Luo, and Yorùbá despite having fewer parameters.

We further analyze the transfer abilities of mBERT and XLM-R by extracting sentence embeddings from the LMs to train a BiLSTM model (*MeanE-BiLSTM*) instead of fine-tuning them end-to-end. Table 5 shows that languages that are not supported by mBERT or XLM-R generally perform worse than CNN-BiLSTM-CRF model (despite being randomly initialized) except for *kin*. Also, sentence embeddings extracted from mBERT often lead to better performance than XLM-R for languages they both do not support (like *ibo*, *kin*, *lug*, *luo*, and *wol*).

Lastly, we train NER models using *language BERT* models that have been adapted to each of the African languages via language-specific fine-tuning on unlabeled text. In all cases, fine-tuning language BERT and language XLM-R models achieves a $1\%-7\%$ improvement in F1-score over fine-tuning mBERT-base and XLM-R-base respectively. This approach is still effective for small sized pre-training corpora provided they are of good quality. For example, the Wolof monolingual corpus, which contains less than 50K sentences (see Table 10 in the Appendix) still improves performance by over 4% F1. Further, we obtain over 60% improvement in performance for Amharic BERT because mBERT does not recognize the Amharic script.

[6] https://github.com/mayhewsw/multilingual-data-stats.

Method	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
CNN-BiLSTM-CRF	50.31	84.64	81.25	60.32	75.66	68.93	62.60	77.83	61.84	66.48	68.99
+ Gazetteers	49.51	**85.02**	80.40	**64.54**	73.85	65.44	**66.54**	**80.16**	**62.44**	65.49	**69.34**

Table 6: Improving NER models using gazetteers. The result is only for 3 Tags: PER, ORG, & LOC. Models trained for 50 epochs. Result is an average over 5 runs.

Method	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
XLM-R-base	69.71	91.03	86.16	73.76	80.51	75.81	86.87	**88.65**	69.56	78.05	77.30
WikiAnn zero-shot	27.68	–	21.90	9.56	–	–	–	36.91	–	10.42	–
eng-CoNLL zero-shot	–	67.52	47.71	38.17	39.45	34.19	67.27	76.40	24.33	39.04	37.15
pcm zero-shot	–	63.71	42.69	40.99	43.50	33.12	–	72.84	25.37	35.16	36.81
swa zero-shot	–	85.35*	55.37	58.44	57.65*	42.88*	72.87*	–	41.70	57.87*	52.32
hau zero-shot	–	–	58.41*	59.10*	59.78	42.81	70.74	83.19*	42.81*	55.97	53.14*
WikiAnn + finetune	**70.92**	–	85.24	72.84	–	–	–	87.90	–	76.78	–
eng-CoNLL + finetune	–	89.73	85.10	71.55	77.34	73.92	84.05	87.59	68.11	75.77	75.30
pcm + finetune	–	90.78	86.42	71.69	79.72	75.56	–	87.62	67.21	78.29	76.48
swa + finetune	–	91.50	87.11	74.84	80.21	74.49	86.74	–	68.47	**80.68**	77.63
hau + finetune	–	–	86.84	74.22	80.56	75.55	88.03	87.92	**70.20**	79.44	77.80
combined East Langs.	–	–	–	**75.65**	81.10	77.56	–	88.15	–	–	–
combined West Langs.	–	90.88	87.06	–	–	–	87.21	–	69.70	**80.68**	–
combined 9 Langs.	–	**91.64**	**87.94**	75.46	**81.29**	**78.12**	**88.12**	88.10	69.84	80.59	78.87

Table 7: Transfer learning result (i.e., F1-score). Three tags: PER, ORG, & LOC. WikiAnn, eng-CoNLL, and the annotated datasets are trained for 50 epochs. Fine-tuning is only for 10 epochs. Results are averaged over 5 runs and the total average (avg) is computed over ibo, kin, lug, luo, wol, and yor languages. The overall highest F1-score is in **bold**, and the best F1-score in zero-shot settings is indicated with an asterisk (*).

6.2 Evaluation of Gazetteer Features

Table 6 shows the performance of the CNN-BiLSTM-CRF model with the addition of gazetteer features as described in Section 5.2.1. On average, the model that uses gazetteer features performs better than the baseline. In general, languages with larger gazetteers, such as Swahili (16K entities in the gazetteer) and Nigerian-Pidgin (for which we use an English gazetteer with 2M entities), have more improvement in performance than those with fewer gazetteer entries, such as Amharic and Luganda (2K and 500 gazetteer entities, respectively). This indicates that having high-coverage gazetteers is important for the model to take advantage of the gazetteer features.

6.3 Transfer Learning Experiments

Table 7 shows the result for the different transfer learning approaches, which we discuss individually in the following sections. We make use of XLM-R-base model for all the experiments in this sub-section because the performance difference if we use XLM-R-large is small (<2%) as shown in Table 5 and because it is faster to train.

Source Language	PER	ORG	LOC
eng-CoNLL	36.17	27.00	50.50
pcm	21.50	65.33	68.17
swa	55.00	69.67	46.00
hau	52.67	57.50	48.50

Table 8: Average per-named entity F1-score for the zero-shot NER using the XLM-R model. The average is computed over ibo, kin, lug, luo, wol, yor languages.

6.3.1 Cross-domain Transfer

We evaluate cross-domain transfer from Wikipedia to the news domain for the five languages that are available in the WikiAnn (Pan et al., 2017) dataset. In the zero-shot setting, the NER F1-score is low: less than 40 F1-score for all

Language	CNN-BiLSTM					mBERT-base					XLM-R-base				
	all	0-freq	0-freq Δ	long	long Δ	all	0-freq	0-freq Δ	long	long Δ	all	0-freq	0-freq Δ	long	long Δ
amh	52.89	40.98	−11.91	45.16	−7.73	–	–	–	–	–	70.96	68.91	−2.05	64.86	−6.10
hau	83.70	78.52	−5.18	66.21	−17.49	87.34	79.41	−7.93	67.67	−19.67	89.44	85.48	−3.96	76.06	−13.38
ibo	78.48	70.57	−7.91	53.93	−24.55	85.11	78.41	−6.70	60.46	−24.65	84.51	77.42	−7.09	59.52	−24.99
kin	64.61	55.89	−8.72	40.00	−24.61	70.98	65.57	−5.41	55.39	−15.59	73.93	66.54	−7.39	54.96	−18.97
lug	74.31	67.99	−6.32	58.33	−15.98	80.56	76.27	−4.29	65.67	−14.89	80.71	73.54	−7.17	63.77	−16.94
luo	66.42	58.93	−7.49	54.17	−12.25	72.65	72.85	0.20	66.67	−5.98	75.14	72.34	−2.80	69.39	−5.75
pcm	66.43	59.73	−6.70	47.80	−18.63	87.78	82.40	−5.38	77.12	−10.66	87.39	83.65	−3.74	74.67	−12.72
swa	79.26	64.74	−14.52	44.78	−34.48	86.37	78.77	−7.60	45.55	−40.82	87.55	80.91	−6.64	53.93	−33.62
wol	60.43	49.03	−11.40	26.92	−33.51	66.10	59.54	−6.56	19.05	−47.05	64.38	57.21	−7.17	38.89	−25.49
yor	67.07	56.33	−10.74	64.52	−2.55	78.64	73.41	−5.23	74.34	−4.30	77.58	72.01	−5.57	76.14	−1.44
avg (excl. amh)	69.36	60.27	−9.09	50.18	−19.18	79.50	74.07	−5.43	59.10	−20.40	79.15	73.80	−5.36	63.22	−15.94

Table 9: F1 score for two varieties of hard-to-identify entities: zero-frequency entities that do not appear in the training corpus, and longer entities of four or more words.

languages, with Kinyarwanda and Yorùbá having less than 10 F1-score. This is likely due to the number of training sentences present in WikiAnn: There are only 100 sentences in the datasets of Amharic, Igbo, Kinyarwanda, and Yorùbá. Although the Swahili corpus has 1,000 sentences, the 35 F1-score shows that transfer is not very effective. In general, cross-domain transfer is a challenging problem, and is even harder when the number of training examples from the source domain is small. Fine-tuning on the in-domain news NER data does not improve over the baseline (XLM-R-base).

6.3.2 Cross-Lingual Transfer

Zero-shot In the zero-shot setting we evaluated NER models trained on the English eng-*CoNLL03* dataset, and on the Nigerian-Pidgin (pcm), Swahili (swa), and Hausa (hau) annotated corpus. We excluded the MISC entity in the eng-*CoNLL03* corpus because it is absent in our target datasets. Table 7 shows the result for the (zero-shot) transfer performance. We observe that the closer the source and target languages are geographically, the better the performance. The pcm model (trained on only 2K sentences) obtains similar transfer performance as the eng-*CoNLL03* model (trained on 14K sentences). swa performs better than pcm and eng-*CoNLL03* with an improvement of over 14 F1 on average. We found that, on average, transferring from Hausa provided the best F1, with an improvement of over 16% and 1% compared to using the eng-*CoNLL* and swa data, respectively. Per-entity analysis in Table 8 shows that the largest improvements are obtained for ORG. The pcm data were more effective in transferring to LOC and ORG, while swa and hau performed better when transferring to PER. In general, zero-shot transfer is most effective when transferring from Hausa and Swahili.

Fine-tuning We use the target language corpus to fine-tune the NER models previously trained on eng-*CoNLL*, pcm, and swa. On average, there is only a small improvement when compared to the XLM-R base model. In particular, we see significant improvement for Hausa, Igbo, Kinyarwanda, Nigerian-Pidgin, Wolof, and Yorùbá using either swa or hau as the source NER model.

6.4 Regional Influence on NER

We evaluate whether combining different language training datasets by region affects the performance for individual languages. Table 7 shows that all languages spoken in West Africa (ibo, wol, pcm, yor) except hau have slightly better performance (0.1–2.6 F1) when we train on their combined training data. However, for the East-African languages, the F1 score only improved (0.8–2.3 F1) for three languages (kin, lug, luo). Training the NER model on all nine languages leads to better performance on all languages except Swahili. On average over six languages (ibo, kin, lug, luo, wol, yor), the performance improves by 1.6 F1.

6.5 Error Analysis

Finally, to better understand the types of entities that were successfully identified and those that were missed, we performed fine-grained analysis of our baseline methods mBERT and XLM-R

using the method of Fu et al. (2020), with results shown in Table 9. Specifically, we found that across all languages, entities that were not contained in the training data (zero-frequency entities), and entities consisting of more than three words (long entities) were particularly difficult in all languages; compared to the F1 score over all entities, the scores dropped by around 5 points when evaluated on zero-frequency entities, and by around 20 points when evaluated on long entities. Future work on low-resource NER or cross-lingual representation learning may further improve on these hard cases.

7 Conclusion and Future Work

We address the NER task for African languages by bringing together a variety of stakeholders to create a high-quality NER dataset for ten African languages. We evaluate multiple state-of-the-art NER models and establish strong baselines. We have released one of our best models that can recognize named entities in ten African languages on HuggingFace Model Hub.[7] We also investigate cross-domain transfer with experiments on five languages with the WikiAnn dataset, along with cross-lingual transfer for low-resource NER using the English CoNLL-2003 dataset and other languages supported by XLM-R. In the future, we plan to use pretrained word embeddings such as GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017) instead of random initialization for the CNN-BiLSTM-CRF, increase the number of annotated sentences per language, and expand the dataset to more African languages.

Acknowledgments

We would like to thank Heng Ji and Ying Lin for providing the ELISA NER tool used for annotation. We also thank the Spoken Language Systems Chair, Dietrich Klakow at Saarland University, for providing GPU resources to train the models. We thank Adhi Kuncoro and the anonymous reviewers for their useful feedback on a draft of this paper. David Adelani acknowledges the support of the EU-funded H2020 project COMPRISE under grant agreement no. 3081705. Finally, we thank Mohamed Ahmed for proofreading the draft.

[7] https://huggingface.co/Davlan/xlm-roberta-large-masakhaner.

References

D. Adelani, Dana Ruiter, J. Alabi, Damilola Adebonojo, Adesina Ayeni, Mofetoluwa Adeyemi, Ayodele Awokoya, and C. España-Bonet. 2021. MENYO-20k: A multi-domain english-yorùbá corpus for machine translation and domain adaptation. *ArXiv*, abs/2103.08647.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: The case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Andrew Caines. 2019. The geographic diversity of NLP conferences.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Guy De Pauw, Peter W Wagacha, and Dorothy Atieno Abade. 2007. Unsupervised induction of Dholuo word classes using maximum entropy learning. *Proceedings of the First International Computer Science and ICT Conference*, page 8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2020. Ethnologue: Languages of the World. 23rd edition.

Roald Eiselen. 2016. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348, Portorož, Slovenia. European Language Resources Association (ELRA).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Nolue Emenanjo. 1978. *Elements of Modern Igbo Grammar - a descriptive approach*. Ibadan, Nigeria. Oxford University Press.

Ignatius Ezeani, Paul Rayson, I. Onyenwe, C. Uchechukwu, and M. Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. *ArXiv*, abs/2004.00648. `https://doi.org/10.1037/h0031619`

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Jinlan Fu, Pengfei Liu, and Graham Neubig. 2020. Interpretable multi-dataset evaluation for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online. Association for Computational Linguistics.

Rwanda Government. 2014. Official gazette number 41 bis of 13/10/2014.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL 2018*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of ICML 2020*.

Zhiheng Huang, W. Xu, and Kailiang Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *ArXiv*, abs/1508.01991.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT 2016*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Ying Lin, Cash Costello, Boliang Zhang, Di Lu, Heng Ji, James Mayfield, and Paul McNamee. 2018. Platforms for non-speakers annotating names in any language. In *Proceedings of ACL 2018, System Demonstrations*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Laura Martinus and Jade Z. Abbott. 2019. A focus on neural machine translation for African languages. *arXiv preprint arXiv:1906.05685*.

MBS. 2020. Téereb Injiil: La Bible Wolof – Ancien Testament. `http://biblewolof.com/`.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online.

Graham Neubig, Chris Dyer, Y. Goldberg, A. Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Manish Kumar, Chaitanya Malaviya, Paul Michel, Y. Oda, M. Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017. Dynet: The dynamic neural network toolkit. *ArXiv*, abs/1701.03980.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Eyo Offiong Mensah. 2012. Grammaticalization in Nigerian Pidgin. *Íkala, revista de lenguaje y cultura*, 17(2):167–179.

Anthony Ojarikre. 2013. Perspectives and problems of codifying Nigerian Pidgin English orthography. *Perspectives*, 3(12).

Ijite Blessing Onovbiona. 2012. Serial verb construction in Nigerian Pidgin.

Ikechukwu E. Onyenwe and Mark Hepple. 2016. Predicting morphologically-complex unknown words in Igbo. In *Text, Speech, and Dialogue*, pages 206–214, Cham. Springer International Publishing.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vuli, Iryna Gurevych, and Sebastian Ruder. 2020a. MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer. In *Proceedings of EMNLP 2020*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. Unks everywhere: Adapting multilingual language models to new scripts. *arXiv preprint arXiv:2012.15562*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using Siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. 2020. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online. Association for Computational Linguistics.

Erik F. Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL 2003*.

Rajeev Sangal, Dipti Misra Sharma, and Anil Kumar Singh. 2008. In *Proceedings of the IJCNLP-08 workshop on named entity recognition for south and south east Asian languages*.

K. Shaalan. 2014. A survey of Arabic named entity recognition and classification. *Computational Linguistics*, 40:469–510. https://doi.org/10.1162/COLI_a_00178

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280. Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art

natural language processing. *ArXiv*, abs/1910.03771.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

A Appendix

A.1 Annotator Agreement

To shed more light on the few cases where annotators disagreed, we provide entity-level confusion matrices across all ten languages in Table 11. The most common disagreement is between organizations and locations.

A.2 Model Hyperparameters for Reproducibility

For fine-tuning mBERT and XLM-R, we used the base and large models with maximum sequence length of 164 for mBERT and 200 for XLM-R, batch size of 32, learning rate of 5e-5, and number of epochs 50. For the MeanE-BiLSTM model, the hyperparameters are similar to fine-tuning the LM except for the learning rate that we set to be 5e-4, the BiLSTM hyperparameters are: input dimension is 768 (since the embedding size from mBERT and XLM-R is 768) in each direction of LSTM, one hidden layer, hidden layer size of 64, and drop-out probability of 0.3 before the last linear layer. All the experiments were performed on a single GPU (Nvidia V100).

A.3 Monolingual Corpora for Language Adaptive Fine-tuning

Table 10 shows the monolingual corpus we used for the language adaptive fine-tuning. We provide the details of the source of the data, and their sizes. For most of the languages, we make use of JW300[8] and CC-100[9]. In some cases CC-Aligned (El-Kishky et al., 2020) was used, in such a case, we removed duplicated sentences from CC-100. For fine-tuning the language model, we make use of the HuggingFace (Wolf et al., 2019) code with learning rate 5e-5. However, for the Amharic BERT, we make use of a smaller learning rate of 5e-6 since the multilingual BERT vocabulary was replaced by Amharic vocabulary, so that we can slowly adapt the mBERT LM to understand Amharic texts. All language BERT models were pre-trained for 3 epochs (‘‘ibo’’, ‘‘kin’’,‘‘lug’’,‘‘luo’’, ‘‘pcm’’,‘‘swa’’,‘‘yor’’) or 10 epochs (‘‘amh’’, ‘‘hau’’,‘‘wol’’) depending on their convergence. The models can be found on HuggingFace Model Hub.[10]

[8] https://opus.nlpl.eu/.
[9] http://data.statmt.org/cc-100/.
[10] https://huggingface.co/Davlan.

Language	Source	Size (MB)	No. sentences
amh	CC-100 (Conneau et al., 2020)	889.7MB	3,124,760
hau	CC-100	318.4MB	3,182,277
ibo	JW300 (Agić and Vulić, 2019), CC-100, CC-Aligned (El-Kishky et al., 2020), and IgboNLP (Ezeani et al., 2020)	118.3MB	1,068,263
kin	JW300, KIRNEWS (Niyongabo et al., 2020), and BBC Gahuza	123.4MB	726,801
lug	JW300, CC-100, and BUKEDDE News	54.0MB	506,523
luo	JW300	12.8MB	160,904
pcm	JW300, and BBC Pidgin	56.9MB	207,532
swa	CC-100	1,800MB	12,664,787
wol	OPUS (Tiedemann, 2012) (excl. CC-Aligned), Wolof Bible (MBS, 2020), and news corpora (Lu Defu Waxu, Saabal, and Wolof Online)	3.8MB	42,621
yor	JW300, Yoruba Embedding Corpus (Alabi et al., 2020), MENYO-20k (Adelani et al., 2021), CC-100, CC-Aligned, and news corpora (BBC Yoruba, Asejere, and Alaroye).	117.6MB	910,628

Table 10: Monolingual corpora, their sources, size, and number of sentences.

	DATE	LOC	ORG	PER
DATE	32,978	–	–	–
LOC	10	70,610	–	–
ORG	0	52	35,336	–
PER	2	48	12	64,216

Table 11: Entity-level confusion matrix between annotators, calculated over all ten languages.