Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2754–2762
Marseille, 11–16 May 2020

Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi

Jesujoba O. Alabi[⋆†‡] **Kwabena Amponsah-Kaakyire**[⋆†‡] **David I. Adelani**[‡‖] **Cristina España-Bonet**[†‡]
[†]DFKI GmbH, Saarbrücken, Germany
[‖]Spoken Language Systems (LSV), Saarland Informatics Campus, [‡]Saarland University, Saarbrücken, Germany
{jesujoba_oluwadara.alabi, kwabena.amponsah-kaakyire, cristinae}@dfki.de, didelani@lsv.uni-saarland.de

Abstract

The success of several architectures to learn semantic representations from unannotated text and the availability of these kind of texts in online multilingual resources such as Wikipedia has facilitated the massive and automatic creation of resources for multiple languages. The evaluation of such resources is usually done for the high-resourced languages, where one has a smorgasbord of tasks and test sets to evaluate on. For low-resourced languages, the evaluation is more difficult and normally ignored, with the hope that the impressive capability of deep learning architectures to learn (multilingual) representations in the high-resourced setting holds in the low-resourced setting too. In this paper we focus on two African languages, Yorùbá and Twi, and compare the word embeddings obtained in this way, with word embeddings obtained from curated corpora and a language-dependent processing. We analyse the noise in the publicly available corpora, collect high quality and noisy data for the two languages and quantify the improvements that depend not only on the amount of data but on the quality too. We also use different architectures that learn word representations both from surface forms and characters to further exploit all the available information which showed to be important for these languages. For the evaluation, we manually translate the wordsim-353 word pairs dataset from English into Yorùbá and Twi. We extend the analysis to contextual word embeddings and evaluate multilingual BERT on a named entity recognition task. For this, we annotate with named entities the Global Voices corpus for Yorùbá. As output of the work, we provide corpora, embeddings and the test suits for both languages.

Keywords: Multilingual embeddings, Low-resource language, Yorùbá, and Twi

1. Introduction

In recent years, word embeddings (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017) have been proven to be very useful for training downstream natural language processing (NLP) tasks. Moreover, contextualized embeddings (Peters et al., 2018; Devlin et al., 2019) have been shown to further improve the performance of NLP tasks such as named entity recognition, question answering, and text classification when used as word features because they are able to resolve ambiguities of word representations when they appear in different contexts. Different deep learning architectures such as multilingual BERT (Devlin et al., 2019), LASER (Artetxe and Schwenk, 2019) and XLM (Lample and Conneau, 2019) have proved successful in the multilingual setting. All these architectures learn the semantic representations from unannotated text, making them *cheap* given the availability of texts in online multilingual resources such as Wikipedia. However, the evaluation of such resources is usually done for the high-resourced languages, where one has a smorgasbord of tasks and test sets to evaluate on. This is the best-case scenario, i.e. languages with tonnes of data for training that generate high-quality models.

For low-resourced languages, the evaluation is more difficult and therefore normally ignored simply because of the lack of resources. In these cases, training data is scarce, and the assumption that the capability of deep learning architectures to learn (multilingual) representations in the high-resourced setting holds in the low-resourced one does not need to be true. In this work, we focus on two African languages, Yorùbá and Twi, and carry out several experiments to verify this claim. Just by a simple inspection of the word embeddings trained on Wikipedia by fastText[1], we see a high number of non-Yorùbá or non-Twi words in the vocabularies. For Twi, the vocabulary has only 935 words, and for Yorùbá we estimate that 135 k out of the 150 k words belong to other languages such as English, French and Arabic.

In order to improve the semantic representations for these languages, we collect online texts and study the influence of the quality and quantity of the data in the final models. We also examine the most appropriate architecture depending on the characteristics of each language. Finally, we translate test sets and annotate corpora to evaluate the performance of both our models together with fastText and BERT pre-trained embeddings which could not be evaluated otherwise for Yorùbá and Twi. The evaluation is carried out in a word similarity and relatedness task using the *wordsim-353* test set, and in a named entity recognition (NER) task where embeddings play a crucial role. Of course, the evaluation of the models in only two tasks is not exhaustive but it is an indication of the quality we can obtain for these two low-resourced languages as compared to others such as English where these evaluations are already available.

The rest of the paper is organized as follows. Related works are reviewed in Section 2. The two languages under study are described in the third section. We introduce the corpora and test sets in Section 4. The fifth section explores the different training architectures we consider, and the experiments that are carried out. Finally, discussion and conclud-

(⋆) Equal contribution to the work, author names are arranged alphabetically by last name.

[1] https://fasttext.cc/docs/en/pretrained-vectors.html

ing remarks are given in Section 6.

2. Related Work

The large amount of freely available text in the internet for multiple languages is facilitating the massive and automatic creation of multilingual resources. The resource par excellence is Wikipedia[2], an online encyclopedia currently available in 307 languages[3]. Other initiatives such as Common Crawl[4] or the Jehovahs Witnesses site[5] are also repositories for multilingual data, usually assumed to be noisier than Wikipedia. Word and contextual embeddings have been pre-trained on these data, so that the resources are nowadays at hand for more than 100 languages. Some examples include fastText word embeddings (Bojanowski et al., 2017; Grave et al., 2018), MUSE embeddings (Lample et al., 2018), BERT multilingual embeddings (Devlin et al., 2019) and LASER sentence embeddings (Artetxe and Schwenk, 2019). In all cases, embeddings are trained either simultaneously for multiple languages, joining high- and low-resource data, or following the same methodology.

On the other hand, different approaches try to specifically design architectures to learn embeddings in a low-resourced setting. Chaudhary et al. (2018) follow a transfer learning approach that uses phonemes, lemmas and morphological tags to transfer the knowledge from related high-resource language into the low-resource one. Jiang et al. (2018) apply Positive-Unlabeled Learning for word embedding calculations, assuming that unobserved pairs of words in a corpus also convey information, and this is specially important for small corpora.

In order to assess the quality of word embeddings, word similarity and relatedness tasks are usually used. *wordsim-353* (Finkelstein et al., 2001) is a collection of 353 pairs annotated with semantic similarity scores in a scale from 0 to 10. Even with the problems detected in this dataset (Camacho-Collados et al., 2017), it is widely used by the community. The test set was originally created for English, but the need for comparison with other languages has motivated several translations/adaptations. In Hassan and Mihalcea (2009) the test was translated manually into Spanish, Romanian and Arabic and the scores were adapted to reflect similarities in the new language. The reported correlation between the English scores and the Spanish ones is 0.86. Later, Joubarne and Inkpen (2011) show indications that the measures of similarity highly correlate across languages. Leviant and Reichart (2015) translated also wordsim-353 into German, Italian and Russian and used crowdsourcing to score the pairs. Finally, Jiang et al. (2018) translated with Google Cloud the test set from English into Czech, Danish and Dutch. In our work, native speakers translate wordsim-353 into Yorùbá and Twi, and similarity scores are kept unless the discrepancy with English is big (see Section 4.2. for details). A similar approach to our work is done for Gujarati in Joshi et al. (2019).

[2] https://www.wikipedia.org

[3] Number of languages in December 2019.

[4] https://commoncrawl.org

[5] https://www.jw.org

3. Languages under Study

Yorùbá is a language in the West Africa with over 50 million speakers. It is spoken among other languages in Nigeria, republic of Togo, Benin Republic and Sierra Leone. It is also a language of Òrìsà in Cuba, Brazil, and some Caribbean countries. It is one of the three major languages in Nigeria and it is regarded as the third most spoken native African language. There are different dialects of Yorùbá in Nigeria (Adegbola, 2016; Asahiah, 2014; Fagbolu et al., 2015). However, in this paper our focus is the standard Yorùbá based upon a report from the 1974 Joint Consultative Committee on Education (Asahiah et al., 2017).

Standard Yorùbá has 25 letters without the Latin characters c, q, v, x and z. There are 18 consonants (b, d, f, g, gb, j[dz], k, l, m, n, p[kp], r, s, ṣ, t, w y[j]), 7 oral vowels (a, e, ẹ, i, o, ọ, u), five nasal vowels, (an, ẹn, in, ọn, un) and syllabic nasals (m̀, ḿ, ǹ, ń). Yorùbá is a tone language which makes heavy use of lexical tones which are indicated by the use of diacritics. There are three tones in Yorùbá namely low, mid and high which are represented as grave (◌̀), macron (◌̄) and acute (◌́) symbols respectively. These tones are applied on vowels and syllabic nasals. Mid tone is usually left unmarked on vowels and every initial or first vowel in a word cannot have a high tone. It is important to note that tone information is needed for correct pronunciation and to have the meaning of a word (Adegbola and Odilinye, 2012; Asahiah, 2014; Asahiah et al., 2017). For example, *owó* (money), *ọwọ̀* (broom), *òwò* (business), *ọ̀wọ̀* (honour), *ọwọ́* (hand), and *ọ̀wọ́* (group) are different words with different dots and diacritic combinations. According to Asahiah (2014), Standard Yorùbá uses 4 diacritics, 3 are for marking tones while the fourth which is the dot below is used to indicate the open phonetic variants of letter "e" and "o" and the long variant of "s". Also, there are 19 single diacritic letters, 3 are marked with dots below (ẹ, ọ, ṣ) while the rest are either having the grave or acute accent. The four double diacritics are divided between the grave and the acute accent as well.

As noted in Asahiah (2014), most of the Yorùbá texts found in websites or public domain repositories (*i*) either use the correct Yorùbá orthography or (*ii*) replace diacritized characters with un-diacritized ones. This happens as a result of many factors, but most especially to the unavailability of appropriate input devices for the accurate application of the diacritical marks (Adegbola, 2016). This has led to research on restoration models for diacritics (Orife, 2018), but the problem is not well solved and we find that most Yorùbá text in the public domain today is not well diacritized. Wikipedia is not an exception.

Twi is an Akan language of the Central Tano Branch of the Niger Congo family of languages. It is the most widely spoken of the about 80 indigenous languages in Ghana (Osam, 2003). It has about 9 million native speakers and about a total of 17–18 million Ghanaians have it as either first or second language. There are two mutually intelligible dialects, Asante and Akuapem, and sub-dialectical variants which are mostly unknown to and unnoticed by non-native speakers. It is also mutually intelligible with Fante and to a large extent Bono, another of the Akan languages.

Description	Source URL	#tokens	Status	C1	C2	C3
Yorùbá						
Lagos-NWU corpus	github.com/Niger-Volta-LTI	24,868	clean	✓	✓	✓
Alákọ̀wé	alakoweyoruba.wordpress.com	24,092	clean	✓	✓	✓
Ọ̀rọ̀ Yorùbá	oroyoruba.blogspot.com	16,232	clean	✓	✓	✓
Èdè Yorùbá Rẹwà	deskgram.cc/edeyorubarewa	4,464	clean	✓	✓	✓
Doctrine $ Covenants	github.com/Niger-Volta-LTI	20,447	clean	✓	✓	✓
Yorùbá Bible	www.bible.com	819,101	clean	✓	✓	✓
GlobalVoices	yo.globalvoices.org	24,617	clean	✓	✓	✓
Jehova Witness	www.jw.org/yo	170,203	clean	✓	✓	✓
Ìrìnkèrindò nínú igbó elégbèje	manual	56,434	clean	✓	✓	✓
Igbó Olódùmarè	manual	62,125	clean	✓	✓	✓
JW300 Yorùbá corpus	opus.nlpl.eu/JW300.php	10,558,055	clean	✗	✗	✓
Yorùbá Tweets	twitter.com/yobamoodua	153,716	clean	✓	✓	✓
BBC Yorùbá	bbc.com/yoruba	330,490	noisy	✗	✓	✓
Voice of Nigeria Yorùbá news	von.gov.ng/yoruba	380,252	noisy	✗	✗	✓
Yorùbá Wikipedia	dumps.wikimedia.org/yowiki	129,075	noisy	✗	✗	✓
Twi						
Bible	www.bible.com	661,229	clean	✓	✓	✓
Jehovah's Witness	www.jw.org/tw	1,847,875	noisy	✗	✗	✓
Wikipedia	dumps.wikimedia.org/twwiki	5,820	noisy	✗	✓	✓
JW300 Twi corpus	opus.nlpl.eu/JW300.php	13,630,514	noisy	✗	✗	✓

Table 1: Summary of the corpora used in the analysis. The last 3 columns indicate in which dataset (C1, C2 or C3) are the different sources included (see text, Section 5.2.).

It is one of, if not the, easiest to learn to *speak* of the indigenous Ghanaian languages. The same is however not true when it comes to *reading* and especially *writing*. This is due to a number of easily overlooked complexities in the structure of the language. First of all, similarly to Yorùbá, Twi is a tonal language but written without diacritics or accents. As a result, words which are pronounced differently and unambiguous in speech tend to be ambiguous in writing. Besides, most of such words fit interchangeably in the same context and some of them can have more than two meanings. A simple example is:

> Me papa aba nti na me ne wo redi no yie no. Sɛ wo ara wo nim sɛ me papa ba a, me suban foforɔ adi.

This sentence could be translated as

> (*i*) I'm only treating you nicely because I'm in a good mood. You already know I'm a completely different person when I'm in a good mood.
>
> (*ii*) I'm only treating you nicely because my dad is around. You already know I'm a completely different person when my dad comes around.

Another characteristic of Twi is the fact that a good number of stop words have the same written form as content words. For instance, "*ɛna*" or "*na*" could be the words "*and, then*", the phrase "*and then*" or the word "*mother*". This kind of ambiguity has consequences in several natural language applications where stop words are removed from text.
Finally, we want to point out that words can also be written with or without prefixes. An example is this same *ɛna* and *na* which happen to be the same word with an omissible prefix across its multiple senses. For some words, the prefix characters are mostly used when the word begins a sentence and omitted in the middle. This however depends on the author/speaker. For the word embeddings calculation, this implies that one would have different embeddings for the same word found in different contexts.

4. Data

We collect *clean* and *noisy* corpora for Yorùbá and Twi in order to quantify the effect of noise on the quality of the embeddings, where noisy has a different meaning depending on the language as it will be explained in the next subsections.

4.1. Training Corpora

For **Yorùbá**, we use several corpora collected by the Niger-Volta Language Technologies Institute[6] with texts from different sources, including the Lagos-NWU conversational speech corpus, fully-diacritized Yorùbá language websites and an online Bible. The largest source with clean data is the JW300 corpus. We also created our own small-sized corpus by web-crawling three Yorùbá language websites (Alàkọ̀wé, Ọ̀rọ̀ Yorùbá and Èdè Yorùbá Rẹwà in Table 1), some Yoruba Tweets with full diacritics and also news corpora (BBC Yorùbá and VON Yorùbá) with poor diacritics which we use to introduce noise. By noisy corpus, we refer to texts with incorrect diacritics (e.g in BBC Yorùbá), removal of tonal symbols (e.g in VON Yorùbá) and removal of all diacritics/under-dots (e.g some articles

[6] https://github.com/Niger-Volta-LTI/yoruba-text

Entity type	Number of tokens			
	Total	Train	Val.	Test
ORG	289	214	40	35
LOC	613	467	47	99
DATE	662	452	86	124
PER	688	469	109	110
O	23,988	17,819	2,413	4,867

Table 2: Number of tokens per named entity type in the Global Voices Yorùbá corpus.

in Yorùbá Wikipedia). Furthermore, we got two manually typed fully-diacritized Yorùbá literature (Ìrìnkèrindò nínú igbó elégbèje and Igbó Olódùmarè) both written by Daniel Orowole Olorunfemi Fagunwa a popular Yorùbá author. The number of tokens available from each source, the link to the original source and the quality of the data is summarised in Table 1.

The gathering of clean data in **Twi** is more difficult. We use the Twi Bible as the base text as it has been shown that the Bible is the most available resource for low-resourced and endangered languages (Resnik et al., 1999). This is the cleanest of all the text we could obtain. In addition, we use the available (and small) Wikipedia dumps which are quite noisy, i.e. Wikipedia contains a good number of English words, spelling errors and Twi sentences formulated in a non-natural way (formulated as L2 speakers would speak Twi as compared to native speakers). Lastly, we added text crawled from Jehovah's Witnesses (2019) and the JW300 Twi corpus. Notice that the Bible text, is mainly written in the Asante dialect whilst the last, Jehovah's Witnesses, was written mainly in the Akuapem dialect. The Wikipedia text is a mixture of the two dialects. This introduces a lot of noise into the embeddings as the spelling of most words differs especially at the end of the words due to the mixture of dialects. The JW300 Twi corpus also contains mixed dialects but is mainly Akuampem. In this case, the noise comes also from spelling errors and the uncommon addition of diacritics which are not standardised on certain vowels. Figures for Twi corpora are summarised in the bottom block of Table 1.

4.2. Evaluation Test Sets

Yorùbá. One of the contribution of this work is the introduction of the wordsim-353 word pairs dataset for Yorùbá. All the 353 word pairs were translated from English to Yorùbá by 3 native speakers. The set is composed of 446 unique English words, 348 of which can be expressed as one-word translation in Yorùbá (e.g. *book* translates to *ìwé*). In 61 cases (most countries and locations but also other content words) translations are transliterations (e.g. *Doctor* is dókítà and *cucumber* is kùkúmbà.). 98 words were translated by short phrases instead of single words. This mostly affects words from science and technology (e.g. *keyboard* translates to pátákó ìtẹ̀wé —literally meaning typing board—, *laboratory* translates to ìyàrá ìṣẹ̀wádìí —research room—, and *ecology* translates to ìmọ̀ nípa àyíká while *psychology* translates to ìmọ̀ nípa ẹ̀dá). Finally, 6 terms have the same form in English and Yorùbá therefore they are retained like that in the dataset (e.g. Jazz, Rock and acronyms such as FBI or OPEC).

We also annotate the Global Voices Yorùbá corpus to test the performance of our trained Yorùbá BERT embeddings on the named entity recognition task. The corpus consists of 26 k tokens which we annotate with four named entity types: DATE, location (LOC), organization (ORG) and personal names (PER). Any other token that does not belong to the four named entities is tagged with "O". The dataset is further split into training (70%), development (10%) and test (20%) partitions. Table 2 shows the number of named entities per type and partition.

Twi Just like Yorùbá, the wordsim-353 word pairs dataset was translated for Twi. Out of the 353 word pairs, 274 were used in this case. The remaining 79 pairs contain words that translate into longer phrases.

The number of words that can be translated by a single token is higher than for Yorùbá. Within the 274 pairs, there are 351 unique English words which translated to 310 unique Twi words. 298 of the 310 Twi words are single word translations, 4 transliterations and 16 are used as is.

Even if Joubarne and Inkpen (2011) showed indications that semantic similarity has a high correlation across languages, different nuances between words are captured differently by languages. For instance, both *money* and *currency* in English translate into *sika* in Twi (and other 32 English words which translate to 14 Twi words belong to this category) and *drink* in English is translated as *Nsa* or *nom* depending on the part of speech (noun for the former, verb for the latter). 17 English words fall into this category. In translating these, we picked the translation that best suits the context (other word in the pair). In two cases, the correlation is not fulfilled at all: *soap–opera* and *star–movies* are not related in the Twi language and the score has been modified accordingly.

5. Semantic Representations

In this section, we describe the architectures used for learning word embeddings for the Twi and Yorùbá languages. Also, we discuss the quality of the embeddings as measured by the correlation with human judgements on the translated wordSim-353 test sets and by the F1 score in a NER task.

5.1. Word Embeddings Architectures

Modeling sub-word units has recently become a popular way to address out-of-vocabulary word problem in NLP especially in word representation learning (Sennrich et al., 2016; Bojanowski et al., 2017; Devlin et al., 2019). A sub-word unit can be a character, character n-grams, or heuristically learned Byte Pair Encodings (BPE) which work very well in practice especially for morphologically rich languages. Here, we consider two word embedding models that make use of character-level information together with word information: Character Word Embedding (CWE) (Chen et al., 2015) and fastText (Bojanowski et al., 2017). Both of them are extensions of the Word2Vec architectures (Mikolov et al., 2013) that model sub-word units, character embeddings in the case of CWE and character n-grams for fastText.

Model	Twi		Yorùbá	
	Vocab Size	**Spearman ρ**	**Vocab Size**	**Spearman ρ**
F1: Pre-trained Model (Wiki)	935	0.143	21,730	0.136
F2: Pre-trained Model (Common Crawl & Wiki)	NA	NA	151,125	0.073
C1: Curated *Small* Dataset (Clean text)	9,923	0.354	12,268	0.322
C2: Curated *Small* Dataset (Clean + some noisy text)	18,494	**0.388**	17,492	0.302
C3: Curated *Large* Dataset (All Clean + Noisy texts)	47,134	0.386	44,560	**0.391**

Table 3: FastText embeddings: Spearman ρ correlation between human judgements and similarity scores on the wordSim-353 for the three datasets analysed (C1, C2 and C3). The comparison with massive fastText embeddings is shown in the top rows.

Model	Twi		Yorùbá	
	Vocab Size	**Spearman ρ**	**Vocab Size**	**Spearman ρ**
C1: Curated *Small* Dataset (Clean text)	21,819	0.377	40,162	0.263
C2: Curated *Small* Dataset (Clean + some noisy text)	22,851	**0.437**	56,086	0.345
C3: Curated *Large* Dataset (All Clean + Noisy texts)	97,913	0.377	133,299	**0.354**

Table 4: CWE embeddings: Spearman ρ correlation between human evaluation and embedding similarities for the three datasets analysed (C1, C2 and C3).

CWE was introduced in 2015 to model the embeddings of characters jointly with words in order to address the issues of character ambiguities and non-compositional words especially in the Chinese language. A word or character embedding is learned in CWE using either CBOW or skipgram architectures, and then the final word embedding is computed by adding the character embeddings to the word itself:

$$x_j = \frac{1}{2}(w_j + \frac{1}{N_j}\sum_{k=1}^{N_j} c_k) \quad (1)$$

where w_j is the word embedding of x_j, N_j is the number of characters in x_j, and c_k is the embedding of the k-th character c_k in x_j.

Similarly, in 2017 fastText was introduced as an extension to skipgram in order to take into account morphology and improve the representation of rare words. In this case the embedding of a word also includes the embeddings of its character n-grams:

$$x_j = \frac{1}{G_j + 1}(w_j + \sum_{k=1}^{G_j} g_k) \quad (2)$$

where w_j is the word embedding of x_j, G_j is the number of character n-grams in x_j and g_k is the embedding of the k-th n-gram.

Chen et al. (2015) also proposed three alternatives to learn multiple embeddings per character and resolve ambiguities: (*i*) position-based character embeddings where each character has different embeddings depending on the position it appears in a word, i.e., beginning, middle or end (*ii*) cluster-based character embeddings where a character can have K different cluster embeddings, and (*iii*) position-based cluster embeddings (CWE-LP) where for each position K different embeddings are learned. We use the latter in our experiments with CWE but no positional embeddings are used with fastText.

Finally, we consider a contextualized embedding architecture, BERT (Devlin et al., 2019). BERT is a masked language model based on the highly efficient and parallelizable Transformer architecture (Vaswani et al., 2017) known to produce very rich contextualized representations for downstream NLP tasks. The architecture is trained by jointly conditioning on both left and right contexts in all the transformer layers using two unsupervised objectives: Masked LM and Next-sentence prediction. The representation of a word is therefore learned according to the context it is found in. Training contextual embeddings needs of huge amounts of corpora which are not available for low-resourced languages such as Yorùbá and Twi. However, Google provided pre-trained multilingual embeddings for 102 languages[7] including Yorùbá (but not Twi).

5.2. Experiments

5.2.1. FastText Training and Evaluation

As a first experiment, we compare the quality of fastText embeddings trained on (high-quality) curated data and

[7] https://github.com/google-research/bert/blob/master/multilingual.md

(low-quality) massively extracted data for Twi and Yorùbá languages.

Facebook released pre-trained word embeddings using fastText for 294 languages trained on Wikipedia (Bojanowski et al., 2017) (identified as F1 in Table 3) and for 157 languages trained on Wikipedia and Common Crawl (Grave et al., 2018) (identified as F2 in Table 3). For Yorùbá, both versions are available but only embeddings trained on Wikipedia are available for Twi. We consider these embeddings the result of training on what we call *massively-extracted corpora*. Notice that training settings for both embeddings are not exactly the same, and differences in performance might come both from corpus size/quality but also from the background model. The 294-languages version is trained using skipgram, in dimension 300, with character n-grams of length 5, a window of size 5 and 5 negatives. The 157-languages version is trained using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives.

We want to compare the performance of these embeddings with the equivalent models that can be obtained by training on the different sources verified by native speakers of Twi and Yorùbá; what we call *curated corpora* and has been described in Section 4. For the comparison, we define 3 datasets according to the quality and quantity of textual data used for training: (*i*) *Curated Small Dataset (clean), C1*, about 1.6 million tokens for Yorùbá and over 735 k tokens for Twi. The clean text for Twi is the Bible and for Yoruba all texts marked under the C1 column in Table 1. (*ii*) In *Curated Small Dataset (clean + noisy), C2*, we add noise to the clean corpus (Wikipedia articles for Twi, and BBC Yorùbá news articles for Yorùbá). This increases the number of training tokens for Twi to 742 k tokens and Yorùbá to about 2 million tokens. (*iii*) *Curated Large Dataset, C3* consists of all available texts we are able to crawl and source out for, either clean or noisy. The addition of JW300 (Agić and Vulić, 2019) texts increases the vocabulary to more than 10 k tokens in both languages.

We train our fastText systems using a skipgram model with an embedding size of 300 dimensions, context window size of 5, 10 negatives and n-grams ranging from 3 to 6 characters similarly to the pre-trained models for both languages. Best results are obtained with minimum word count of 3. Table 3 shows the Spearman correlation between human judgements and cosine similarity scores on the wordSim-353 test set. Notice that pre-trained embeddings on Wikipedia show a very low correlation with humans on the similarity task for both languages (ρ=0.14) and their performance is even lower when Common Crawl is also considered (ρ=0.07 for Yorùbá). An important reason for the low performance is the limited vocabulary. The pre-trained Twi model has only 935 tokens. For Yorùbá, things are apparently better with more than 150 k tokens when both Wikipedia and Common Crawl are used but correlation is even lower. An inspection[8] of the pre-trained embeddings indicates that over 135 k words belong to other languages mostly English, French and Arabic. If we focus only on Wikipedia, we see that many texts are without diacritics in Yorùbá and often make use of mixed dialects and English sentences in Twi.

The Spearman ρ correlation for fastText models on the curated small dataset (clean), C1, improves the baselines by a large margin ($\rho = 0.354$ for Twi and 0.322 for Yorùbá) even with a small dataset. The improvement could be justified just by the larger vocabulary in Twi, but in the case of Yorùbá the enhancement is there with almost half of the vocabulary size. We found out that adding some noisy texts (C2 dataset) slightly improves the correlation for Twi language but not for the Yorùbá language. The Twi language benefits from Wikipedia articles because its inclusion doubles the vocabulary and reduces the bias of the model towards religious texts. However, for Yorùbá, noisy texts often ignore diacritics or tonal marks which increases the vocabulary size at the cost of an increment in the ambiguity too. As a result, the correlation is slightly hurt. One would expect that training with more data would improve the quality of the embeddings, but we found out with the results obtained with the C3 dataset, that only high-quality data helps. The addition of JW300 boosts the vocabulary in both cases, but whereas for Twi the corpus mixes dialects and is noisy, for Yorùbá it is very clean and with full diacritics. Consequently, the best embeddings for Yorùbá are obtained when training with the C3 dataset, whereas for Twi, C2 is the best option. In both cases, the curated embeddings improve the correlation with human judgements on the similarity task a $\Delta\rho = +0.25$ or, equivalently, by an increment on ρ of 170% (Twi) and 180% (Yorùbá).

5.2.2. CWE Training and Evaluation

The huge ambiguity in the written Twi language motivates the exploration of different approaches to word embedding estimations. In this work, we compare the standard fastText methodology to include sub-word information with the character-enhanced approach with position-based clustered embeddings (CWE-LP as introduced in Section 5.1.). With the latter, we expect to specifically address the ambiguity present in a language that does not translate the different oral tones on vowels into the written language.

The character-enhanced word embeddings are trained using a skipgram architecture with cluster-based embeddings and an embedding size of 300 dimensions, context window-size of 5, and 5 negative samples. In this case, the best performance is obtained with a minimum word count of 1, and that increases the effective vocabulary that is used for training the embeddings with respect to the fastText experiments reported in Table 3.

We repeat the same experiments as with fastText and summarise them in Table 4. If we compare the relative numbers for the three datasets (C1, C2 and C3) we observe the same trends as before: the performance of the embeddings in the similarity task improves with the vocabulary size when the training data can be considered clean, but the performance diminishes when the data is noisy.

According to the results, CWE is specially beneficial for Twi but not always for Yorùbá. Clean Yorùbá text, does not have the ambiguity issues at character-level, therefore the n-gram approximation works better when enough clean

[8] We used *langdetect* to have a rough estimation of the language of each word, assuming that words that are not detected are Yorùbá because the language is not supported by the tool.

Embedding Type	DATE	LOC	ORG	PER	F1-score
Pre-trained *uncased* Multilingual-bert (Multilingual vocab)	44.6	33.9	12.1	5.7	27.1 ± 0.7
Fine-tuned *uncased* Multilingual-bert (Multilingual vocab)	64.0	65.3	38.8	47.4	56.4 ± 2.4
Fine-tuned *uncased* Multilingual-bert (Yorùbá vocab)	67.0	71.5	40.4	49.4	60.1 ± 0.8

Table 5: NER F1 score on Global Voices Yorùbá corpus after fine-tuning BERT for 10 epochs. Mean F1-score computed after 5 runs

data is used ($\rho^{C3}_{CWE} = 0.354$ vs. $\rho^{C3}_{fastText} = 0.391$) but it does not when too much noisy data (no diacritics, therefore character-level information would be needed) is used ($\rho^{C2}_{CWE} = 0.345$ vs. $\rho^{C2}_{fastText} = 0.302$). For Twi, the character-level information reinforces the benefits of clean data and the best correlation with human judgements is reached with CWE embeddings ($\rho^{C2}_{CWE} = 0.437$ vs. $\rho^{C2}_{fastText} = 0.388$).

5.2.3. BERT Evaluation on NER Task

In order to go beyond the similarity task using static word vectors, we also investigate the quality of the multilingual BERT embeddings by fine-tuning a named entity recognition task on the Yorùbá Global Voices corpus.

One of the major advantages of pre-trained BERT embeddings is that fine-tuning of the model on downstream NLP tasks is typically computationally inexpensive, often with few number of epochs. However, the data the embeddings are trained on has the same limitations as that used in massive word embeddings. Fine-tuning involves replacing the last layer of BERT used optimizing the masked LM with a task-dependent linear classifier or any other deep learning architecture, and training all the model parameters end-to-end. For the NER task, we obtain the token-level representation from BERT and train a conditional random field classifier for sequence tagging.

Similar to our observations with non-contextualized embeddings, we find out that fine-tuning the pre-trained multilingual-uncased BERT for 10 epochs on the NER task gives an F1 score of 27. If we do the same experiment in English, F1 is 66.2 after 10 epochs. That shows how pre-trained embeddings by themselves do not perform well in downstream tasks on low-resource languages. To address this problem for Yorùbá, we fine-tune BERT masked language model on the Yorùbá corpus in two ways: (*i*) using the multilingual vocabulary, and (*ii*) using only Yorùbá vocabulary. In both cases diacritics are ignored to be consistent with the base model training.

As expected, the fine-tuning of the pre-trained BERT on the Yorùbá corpus in the two configurations generates better representations than the base model. These models are able to achieve a better performance on the NER task with an average F1 score of over 56% (see Table 5 for the comparative). The fine-tuned BERT model with only Yorùbá vocabulary further increases by 4% in F1 score than the BERT model that uses the multilingual vocabulary. Although we do not have enough data to train BERT from scratch, we observe that fine-tuning BERT on a limited amount of monolingual data of a low-resource language helps to improve the quality of the embeddings. The same observation holds true for high-resource languages like German[9] and French (Martin et al., 2019).

6. Summary and Discussion

In this paper, we present curated word and contextual embeddings for Yorùbá and Twi. For this purpose, we gather and select corpora and study the most appropriate techniques for the languages. We also create test sets for the evaluation of the word embeddings within a word similarity task (wordsim353) and the contextual embeddings within a NER task. Corpora, embeddings and test sets are available in github[10].

In our analysis, we show how massively generated embeddings perform poorly for low-resourced languages as compared to the performance for high-resourced ones. This is due both to the quantity but also the quality of the data used. While the Pearson ρ correlation for English obtained with fastText embeddings trained on Wikipedia (WP) and Common Crawl (CC) are ρ_{WP}=0.67 and ρ_{WP+CC}=0.78, the equivalent ones for Yorùbá are ρ_{WP}=0.14 and ρ_{WP+CC}=0.07. For Twi, only embeddings with Wikipedia are available (ρ_{WP}=0.14). By carefully gathering high-quality data and optimising the models to the characteristics of each language, we deliver embeddings with correlations of ρ=0.39 (Yorùbá) and ρ=0.44 (Twi) on the same test set, still far from the high-resourced models, but representing an improvement over 170% on the task.

In a low-resourced setting, the data quality, processing and model selection is more critical than in a high-resourced scenario. We show how the characteristics of a language (such as diacritization in our case) should be taken into account in order to choose the relevant data and model to use. As an example, Twi word embeddings are significantly better when training on 742 k selected tokens than on 16 million noisy tokens, and when using a model that takes into account single character information (CWE-LP) instead of n-gram information (fastText).

Finally, we want to note that, even within a corpus, the quality of the data might depend on the language. Wikipedia is usually used as a high-quality freely available multilingual corpus as compared to noisier data such as Common Crawl. However, for the two languages under study, Wikipedia resulted to have too much noise: interference from other languages, text clearly written by non-native speakers, lack of diacritics and mixture of dialects. The JW300 corpus on the other hand, has been rated as high-quality by our native Yorùbá speakers, but as noisy by our native Twi speakers. In both cases, experiments confirm the conclusions.

7. Acknowledgements

The authors thank Dr. Clement Odoje of the Department of Linguistics and African Languages, University of Ibadan,

[9] https://deepset.ai/german-bert

[10] https://github.com/ajesujoba/YorubaTwi-Embedding

Nigeria and Olóyè Gbémisóyè Àrẹ̀ó for helping us with the Yorùbá translation of the WordSim-353 word pairs and Dr. Felix Y. Adu-Gyamfi and Ps. Isaac Sarfo for helping with the Twi translation. We also thank the members of the Niger-Volta Language Technologies Institute for providing us with clean Yorùbá corpus

The project on which this paper is based was partially funded by the German Federal Ministry of Education and Research under the funding code 01IW17001 (Deeplee). Responsibility for the content of this publication is with the authors.

8. Bibliographical References

Adegbola, T. and Odilinye, L. U. (2012). Quantifying the effect of corpus size on the quality of automatic diacritization of yoruba texts. In *Spoken Language Technologies for Under-Resourced Languages*.

Adegbola, T. (2016). Pattern-based unsupervised induction of yoruba morphology. In *Proceedings of WWW 2016 Companion*, pages 599–604.

Agić, Ž. and Vulić, I. (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, July. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. volume 7, pages 597–610. MIT Press, September.

Asahiah, F. O., Odejobi, O. A., and Adagunodo, E. R. (2017). Restoring tone-marks in standard yoruba electronic text: Improved model. *Computer Science*, 18(3):301–315.

Asahiah, F. O. (2014). Development of a standard yoruba digital text automatic diacritic restoration system. *Phd. Thesis, Obafemi Awolowo University, Ile-Ife, Nigeria*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Camacho-Collados, J., Pilehvar, M. T., Collier, N., and Navigli, R. (2017). SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26, Vancouver, Canada, August. Association for Computational Linguistics.

Chaudhary, A., Zhou, C., Levin, L. S., Neubig, G., Mortensen, D. R., and Carbonell, J. G. (2018). Adapting word embeddings to new languages with morphological and phonological subword representations. In Ellen Riloff, et al., editors, *EMNLP*, pages 3285–3295. Association for Computational Linguistics.

Chen, X., Xui, L., Zhiyuan, L., Sun, M., and Luan, H. (2015). Joint learning of character and word embeddings. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1236–1242. IJCAI.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Fagbolu, O., Ojoawo, A., Ajibade, K., and Alese, B. (2015). Digital yoruba corpus. *International Journal of Innovative Science, Engineering Technology*, 2(8):918–926.

Finkelstein, I., Gabrilovich, E., Mathias, Y., Rivlin, E., Solan, Z., and Wolfman, G. (2001). Placing search in context: The concept revisited. In *10th International Conference on World Wide Web*, pages 406–414.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hassan, S. and Mihalcea, R. (2009). Cross-lingual semantic relatedness using encyclopedic knowledge. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1201, Singapore, August. Association for Computational Linguistics.

Jehovah's Witnesses. (2019).

Jiang, C., Yu, H.-F., Hsieh, C.-J., and Chang, K.-W. (2018). Learning word embeddings for low-resource languages by PU learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1024–1034, New Orleans, Louisiana, June. Association for Computational Linguistics.

Joshi, I., Koringa, P., and Mitra, S. (2019). Word embeddings in low resource gujarati language. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 5, pages 110–115, Sep.

Joubarne, C. and Inkpen, D. (2011). Comparison of semantic similarity for different languages using the google n-gram corpus and second-order co-occurrence measures. In Cory Butz et al., editors, *Advances in Artificial Intelligence*, pages 216–221, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Lample, G., Conneau, A., Ranzato, M., Denoyer, L., and Jégou, H. (2018). Word translation without parallel data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Leviant, I. and Reichart, R. (2015). Judgment language matters: Multilingual vector space models for judgment language aware lexical semantics. *CoRR*, abs/1508.00106.

Martin, L., Muller, B., Surez, P. J. O., Dupont, Y., Romary, L., ric Villemonte de la Clergerie, Seddah, D., and Sagot, B. (2019). Camembert: a tasty french language model.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing System*, pages 3111–3119.

Orife, I. (2018). Attentive sequence-to-sequence learning for diacritic restoration of yorùbá language text. *Proc. Interspeech 2018*, pages 2848–2852.

Osam, E. K. (2003). An introduction to the verbal and multi-verbal system of akan. In *Proceedings of the workshop on Multi-Verb Constructions*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Conference Proceedings - Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Resnik, P., Olsen, M. B., and Diab, M. T. (1999). The bible as a parallel corpus: Annotating the book of 2000 tongues. *Computers and the Humanities*, 33:129–153.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.