AfriBERTa: Towards Viable Multilingual Language Models for Low-resource Languages

by

Kelechi Ogueji

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

There are over 7000 languages spoken on earth, but many of these languages suffer from a dearth of natural language processing (NLP) tools. Multilingual pretrained language models have been introduced to help alleviate this problem. However, the largest pretrained multilingual models were trained on only hundreds of languages. This is a small amount when compared to the number of spoken languages. While these models have displayed impressive performance on several languages, including those they were not pretrained on, there is a lot of ground to be covered.

A lot of languages are often left out because pretrained language models are assumed to require a lot of training data, which the languages do not have. Furthermore, a major motivation behind these models is that such lower-resource languages benefit from joint training with higher-resource languages. In this thesis, we challenge both these assumptions and present the first attempt at training a multilingual language model on only low-resource languages. We show that it is possible to train competitive multilingual language models on less than one gigabyte of text data containing a selection of African languages.

Our model, named AfriBERTa, covers 11 African languages, including the first language model for 4 of these languages. We evaluate this model on named entity recognition and text classification spanning 10 languages. Our evaluation results show that our model is very competitive with larger multilingual models - multilingual BERT and XLM-RoBERTa - on several languages. Results suggest that our "small data" approach based on similar languages may sometimes work better than joint training on large datasets with high-resource languages. Furthermore, we present a comprehensive discussion of the implications of our findings.

Acknowledgements

I would like to start off by thanking Professor Jimmy Lin for his impeccable guidance and support during the course of my program. I would like to especially thank him for allowing a focus on natural language processing research for African languages.

Also, my sincere gratitude goes to the readers of my thesis, Professors Charles Clarke and Professor Mei Nagappan for taking the time out to review my thesis.

Furthermore, I would like to acknowledge members of the Data Systems Group (DSG) for their insightful discussions and help, especially Crystina and Rodrigo.

Finally, my special thanks goes to my family for their unwavering love and support all through these years.

Dedication

This is dedicated to God and my family for their immense love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

A lot of recent progress in natural language processing (NLP) has been achieved by the use of neural network architectures. Recurrent neural networks (RNN) [58], long short term memories (LSTM) [26], gated recurrent units (GRU) [14] are examples of such architectures. More recently, the transformer architecture was introduced [59]. The transformer is based on an attention mechanism and explicitly models the relationship between tokens in a sequence. A synthesis of this architecture and self-supervised learning has birthed pretrained language models (PLM). In this setting, a transformer-based architecture is trained (pretrained) in a self-supervised manner on very large corpora, learning general representations. These representations can then be used to aid downstream NLP tasks, such as text classification, by training (finetuning) the pretrained language models on labelled data of such tasks. Examples of these models include BERT [18], RoBERTa [35], XLNet [65] and T5 [50].

Despite the fact that these models have proven to be the de-facto method for a lot of NLP tasks because of their effectiveness, it is expensive and often impractical to train a single pretrained language models for every single language. Hence, pretrained language models have been extended to the multilingual setting. In this setting, a single model is pretrained on a concatenation of text corpora from several languages. Such models have been shown to possess cross-lingual capabilities across many languages. Examples of these models include XLM-R [15], mBERT [18] and mT5 [64].

For all their promise, these models are known to require a lot of training data [1], which is absent for many languages. This consequently leaves out many of the over 7000 languages on earth from these models. Languages with little to no training corpora are commonly described as low-resource in NLP, while those with abundant corpora are described as

high-resource. High-resource languages usually make up a significant part of the training data for multilingual pretrained language models, as it is hypothesized that they help boost the performance of lower-resource languages via cross-lingual transfer. Hence, there has been no previous attempt to investigate if it is possible to pretrain multilingual language models solely on low-resource languages without any transfer from higher-resource languages, despite the numerous benefits that this could provide, some of which are discussed in section 5.3.

In this thesis, we describe our work [44] which aims to cover this gap in the literature. The goal of this work is to explore the viability of multilingual language models pretrained from scratch on low-resource languages and to understand how to pretrain such models in this setting. We introduce AfriBERTa, a family of transformer-based multilingual language models trained on 11 African languages, all of which are low-resource.[1] We evaluate this model on named entity recognition (NER) and text classification downstream tasks on 10 low-resource languages. Our models outperform larger models like mBERT and XLM-R by up to 10 F1 points on text classification, and also outperform these models on several languages in the NER task. Across all languages, we obtain very competitive performance to these larger models. Our results show that, for the first time, it is possible to pretrain a multilingual language model from scratch on only low-resource languages and obtain good performance on downstream tasks.

1.1 Contributions

In summary, our contributions are as follows:

1. We introduce the first pretrained language models for 4 African languages, improving the representation of low-resource languages in modern NLP tools.
2. Using a case study on African languages, we show that competitive multilingual language models can be pretrained from scratch solely on low-resource languages without any high-resource transfer.
3. We show that it is possible to pretrain these models on less than one gigabyte of text data from a selection of African languages, and highlight the many practical benefits of this.

[1] One of the languages (Gahuza) is counted twice because it is a code-mixed language consisting of Kinyarwanda and Kirundi.

4. We release the corpora to the community so as to stimulate future research on African languages.

5. Our extensive experiments highlight important factors to consider when pretraining multilingual language models in low-resource settings. For example, we find that increasing the vocabulary size does not always yield better results when pretraining on smaller datasets. While a small vocabulary size performs relatively poorly, medium sized vocabularies can sometimes outperform larger ones. This is the opposite of what has been found for larger datasets [35].

1.2 Thesis Organization

The thesis is organized as follows:

1. In Chapter 2, we cover the related work and background knowledge required to understand our work.

2. Chapter 3 introduces the proposed approach of our work, highlighting the workings and model objective of AfriBERTa.

3. Chapter 4 describes our experimental setup, tasks, datasets, and languages covered by our work. We also exhaustively discuss the implementation details of this setup.

4. Chapter 5 discusses our results. Furthermore, we provide an in-depth discussion of the implications of those results.

5. Chapter 6 concludes the thesis by summarizing the main contributions and highlighting future work.

Chapter 2

Background and Related Work

2.1 Distributed Representation of Words

Natural language text needs to be represented in numeric form in order for computer systems to be able to process them. This has been an important focus area of natural language processing (NLP) research. Traditional term-based methods such as TF-IDF do not capture the semantic meaning of words and have several drawbacks. Subsequently, methods based on the distributional hypothesis [23, 24] - which states that words that often appear in the same contexts are likely to have similar meanings - were introduced.

Word2Vec [41] produces distributed word representations (real-valued vectors) by using a shallow neural network (usually a single hidden layer) and training it via self-supervised learning. The model can be trained via two settings: the continuous bag of words (CBOW) method where a word is predicted based on its context (words that precede it) and the skip-gram method where contexts are predicted based on an input word.

GloVe [46] is another method that learns word representations in an unsupervised manner. This is achieved by combining the global corpus statitics with local window contexts. GloVe has been shown to outperform Word2Vec on several benchmarks [46].

The methods described above learn representations for whole words; however, there is a lot of benefit in learning representations for subwords. For example, subword models have been shown to perform better on morphologically rich languages and rare words, in comparison to word-level models. Fasttext [12] learns representations for subwords (character n-grams) via CBOW and skip-gram methods described previously. To obtain the representation of a word, a bag of character n-grams is used.

While the models described above showed good performances on a plethora of natural language processing tasks, they all share the major drawback of being context-independent. Human language is very contextual, and we need to convey this to computer systems. For example, consider the two sentences below:

1. I am eating an apple.
2. Apple just released a new phone.

As a human, one can infer that the "apple" in the first sentence refers to the fruit, while the "apple" in the second sentence refers to the technology company. However, if we used any of the models described above, we would get the same representation for the word "apple". Hence, we need a way to obtain a representation for a word, depending on the context it is in.

ELMo [47] (Embeddings from Language Models) is an unsupervised model that produces contextual text representations. This means that polysemy of words can be accounted for, and in different contexts, the same word can have different vector representations. ELMo uses a Bidirectional Long Short Term Memory (BiLSTM) model to capture contexts in both directions. The word representations are obtained from the internal states of the BiLSTM.

2.2 Transformer-based Contextual Encoders

2.2.1 Monolingual Models

The introduction of transformers [59] has advanced several natural language processing tasks, including learning unsupervised text representations. Transformers consist of layers of multiple self-attention heads which aim to capture the importance of tokens in an input sequence. For every token in an input sequence, an attention head computes key, query and value vectors which are used in calculating a weighted representation. The resulting output from all the heads in a layer are combined and fed into a full-connected layer. Skip connections and layer normalization are also used in each layer. The original transformer is a sequence-to-sequence model, meaning it contains both an encoder and decoder, all based on the multi-headed self-attention mechanism.

Bidirectional Encoder Representation from Transformers (BERT) [18] is a transformer-based model which can learn contextual representations and has been shown to significantly

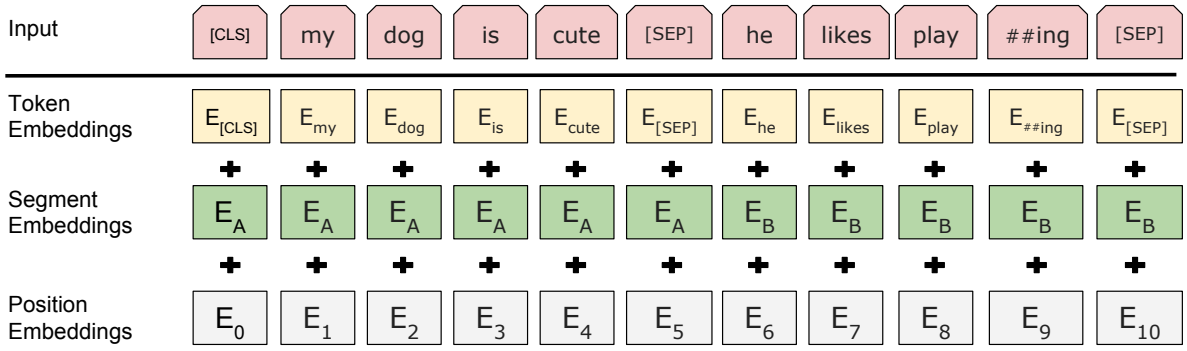

Figure 2.1: BERT input embeddings: The input embeddings are the sum of the token embeddings, the segment embeddings and the position embeddings. Adapted from Delvin et al. [18].

outperform all previous methods discussed on tasks such as sentiment analysis, named entity recognition and question answering. BERT is essentially a stack of transformer encoder layers. In order to make BERT usable for several tasks, its input format is carefully designed to be generic. A special token, [CLS], is always placed at the start of an input sequence. There exists another special token, [SEP], which is used as a delimiter if an input sequence contains more than one sentence, such as question answering or natural language inference tasks. For tokenization, BERT uses a WordPiece model [63] with a vocabulary of 30,000 tokens. The WordPiece model helps divide words into smaller subwords. In order to get a token embedding, the subword embedding is added to a positional embedding (which helps denote order) and a segment embedding (which helps denote the input sentence the token belongs to). Details of the input embedding are illustrated in Figure 2.1.

The pretraining task introduced in BERT is perhaps its biggest novelty. BERT is pretrained using masked language model (MLM) and next sentence prediction (NSP) objectives. In the MLM task, 15% of the input tokens are randomly selected to be masked with a [MASK] token and the model is tasked to predict what tokens are masked. Since the [MASK] token is never seen during finetuning, the authors propose a variation of total masking. In the variant used, 15% of the token positions are selected at random and 80% of these positions are replaced with the mask token, 10% are replaced with a random token and 10% of the tokens are left unchanged. The model is then tasked with predicting the original token with a cross entropy loss. In the NSP task, two sentences are inputted into the model and the model is charged with predicting whether or not they are adjascent sentences in the pretraining corpus. 50% of the time, both sentences are adjascent to each other, and 50% of time, they are not. This training objective helps the model learn the

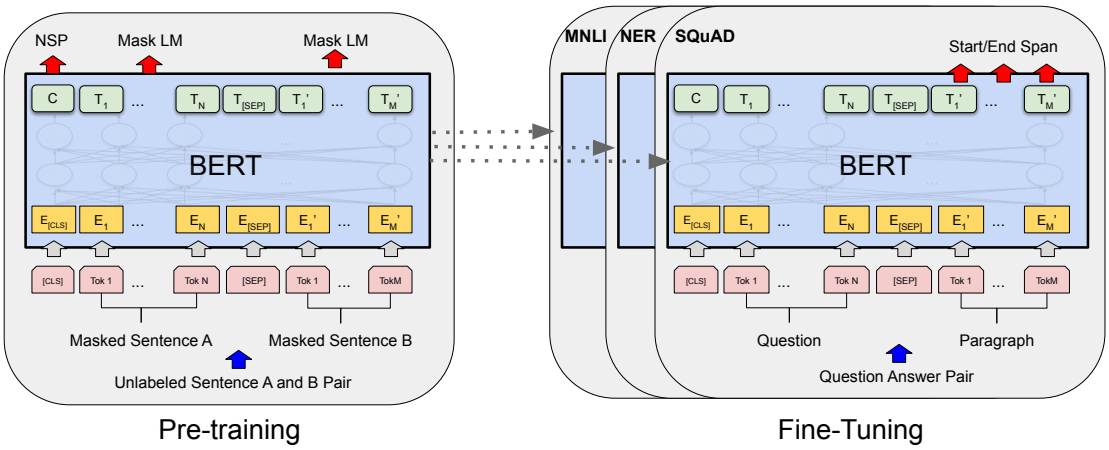

Figure 2.2: General pretraining and finetuning procedures for BERT. The same architectures are used for both pretraining and finetuning stages, save for the output layers. The model parameters from pretraining are used to initialize the models for various downstream tasks. During finetuning, all parameters may be finetuned. The [CLS] token is a special token that is prepended to the front of every input sequence, and [SEP] is also a special token that separates input sentences. Adapted from Delvin et al. [18].

relationship between sentencces and has been shown to help multi-sentence NLP tasks such as question answering and natural language inference. BERT is pretrained on a combination of the Google BookCorpus containing about 800 million words [67] and the English Wikipedia containing over 2 billion words.

Pretraining BERT allows it amass a lot of semantic and syntactic knowledge [52] which can be transferred to downstream NLP tasks, such as text classification and token classification. The knowledge from BERT can be transferred to downstream tasks by finetuning the pretrained model by appending a task-specific layer to its final hidden layer. For token-level tasks such as named entity recognition, the output of each token is passed into the task-specific layers. While for sequence-level tasks, such as sequence classification, the output representation of the [CLS] special token is used as the entire sequence representation and is passed into the task-specific layers. The parameters of the addition layer and BERT are then both finetuned to maximize the log probability of the correct label. Figure 2.2 shows the pretraining and finetuning procedures for BERT.

There have been other variants of BERT which proposed different pretraining objectives, such as removing the NSP [35] and span-level corruption [27]. Furthermore, while the original BERT was trained on only English language, there have been extensions to

other languages with largely successful results [38, 48, 53, 17].

2.2.2 Multilingual Models

Given the several thousands of spoken languages, it is quite an ask to train a BERT model for each and every one of these languages. Hence, several works attempted to kill several birds with one stone by training a single BERT model for many languages. A multilingual version of BERT (mBERT) [18] was released by its original authors, covering 104 languages.[1] The languages selected were those with the largest Wikipedia data. For each language, its entire Wikipedia dump (excluding user and talk pages) was used as the training data. Since different languages have varying amounts of training data, it is easy to overfit or underfit on certain languages. Hence, an exponentially smoothed weighting method is used to undersample languages with a lot of data and oversample those with little data. This weighting is also used for the data used in training the tokenizer. Just like in monolingual BERT, WordPiece tokenizer [54] is used but with a larger vocabulary size of 110k to accommodate the increased number of languages. When an input sequence is fed into the model, there is no mark denoting the language it is from. There is also no explicit cross-lingual supervision used during training, enabling the model to learn the association between the languages all by itself.

XLM-RoBERTa (XLM-R) [15] improved upon mBERT by training a larger model on more data. The authors also exposed a trade-off as the number of languages are increased for a fixed model capacity, which they refer to as the *curse of multilinguality*. Their model is trained on 100 languages with data obtained from Common Crawl [61]. They showed that the trade-off previously described can be alleviated by increasing model and vocabulary size. They also showed that, in general, longer training time and larger scale data benefited their models.

A major hypothesis of both models above is that high-resourced languages can help low-resourced via cross-lingual transfer. While this has been shown to be beneficial, it casts an implicit assumption that low-resourced language models cannot be successfully trained without this transfer.

2.2.3 Pretraining Contextual Encoders with Small Data

Pretrained language models have been shown to perform well when there is a lot of data [35, 15], but some works have focused on using relatively smaller amounts of data. Camem-

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

BERT [37] showed that it is possible to obtain state-of-the-art result with a French BERT model pretrained on small-scale diverse data. In another work [40], the authors showed that training a French BERT language model on 100 MB of data yields similar performance on question answering as models pretrained on larger datasets. Furthermore, state-of-the-art performance has been obtained with ELMo [47] language models pretrained on less than 1 GB of Wikipedia text [45]. It has also been shown that RoBERTa language models [35] trained on 10 to 100 million tokens can encode most syntactic and semantic features in its learned text representations [66].

A common theme among these works is their focus on monolingual language models. While it is possible to learn monolingual language models on smaller amounts of data, it remains to be seen if it is possible in the multilingual case. Our work is the first, to the best of our knowledge, that focuses on pretraining a multilingual language model solely on low-resource languages without any transfer from higher-resource languages.

2.3 Language Representation in NLP

Despite interesting progress in both monolingual and multilingual pretrained models, much of this progress has been focused on languages with relatively large amounts of data, commonly referred to as *high-resource languages*. There has especially been very little focus on African languages, despite the over 2000 languages spoken on the continent making up 30.1% of all living languages [20]. This is further visible in NLP publications on these languages. In all the Association for Computational Linguistics (ACL) conferences hosted in 2019, only 0.19% author affiliations were located in Africa [13]. Other works [28] have also noted the great disparity in the coverage of languages by NLP technologies. They note that over 90% of the world's 7000+ languages are under-studied by the NLP community.

There have been a few works on learning pretrained embeddings for African languages, although many of them have been static and trained on a specific language [21, 43, 6, 19]. More recently, Azunre et al. [8] trained a BERT model on the Twi language. However, they note that their model is biased to the religious domain because much of their data comes from that domain.

While some African languages have been included in multilingual language models, this coverage only scratches the surface of the number of spoken African languages. Furthermore, the languages always make up a minuscule percentage of the training set. For instance, amongst the 104 languages that mBERT was pretrained on, only 3 are African. In XLM-R, there are only 8 African languages out of the 100 languages. In terms of

dataset size, the story is the same. African languages make up 4.80 GB out of about 2395 GB that XLM-R was pretrained on, representing just 0.2% of the entire dataset [15]. In mBERT, African languages make up just 0.24 GB out of the approximately 100 GB that the model was pretrained on. All of this call for an obvious need for increased representation of African languages in modern NLP tools for the over 1.3 billion speakers on the continent.[2]

[2] https://www.worldometers.info/world-population/africa-population/

Chapter 3

Proposed Approach

In this chapter, we discuss our approach to learning multilingual language models for low-resourced African languages. While our approach largely follows well-established methods of learning multilingual masked language models as described in multilingual BERT [18] and XLM-R [15], we elaborate more on the model mechanisms and discuss in detail some subtle differences that makes this approach work on our small-sized dataset.

3.1 Multilingual Masked Language Model Pretraining Objective

Using a standard transformer architecture, we perform masked language modelling (MLM) where 15% of the input tokens are randomly selected to be masked with a special mask token and the model is tasked to predict what tokens are masked. Specifically, 15% of the token positions are selected at random and 80% of these positions are replaced with the mask token, 10% are replaced with a random token and 10% of the tokens are left unchanged. We do not use the next sentence prediction task that was used in Delvin et al. [18] nor do we use the translation language modelling task that was used in Lample and Conneau [16]. Instead, we use only the (MLM) approach following Conneau et al. [15]. Figure 3.1 illustrates the MLM objective described.

Since we want our model to be multilingual, the batches fed into our model come from different languages. Unlike Delvin et al. [18] where different languages can be in one batch, we ensure that a single batch contains the same language as preliminary results showed that this performed better in our small-data regime.

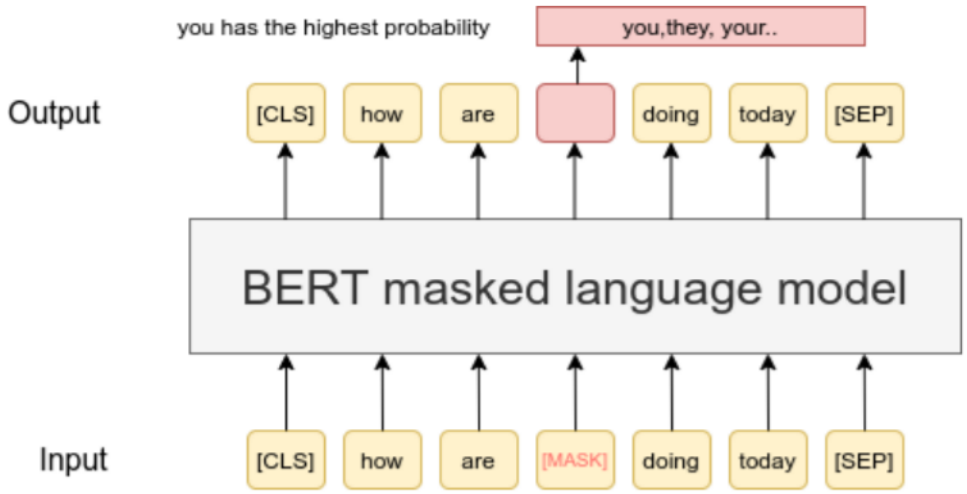

Figure 3.1: Masked language model pretraining mechanism: The model is tasked with predicting the masked out tokens. Figure taken from Reimers and Gurevych, 2019 [51]

3.2 Training the Tokenizer

We learn a shared vocabulary for all languages as this signficantly contributes to the alignment of the embedding spaces across all languages. We utilize subword tokenization on the raw text data using SentencePiece [32] trained with a unigram language model [31]. Our models are trained on N languages. Hence, we have N monolingual corpora $\{D_i\}_{i=1...N}$, and we denote by n_i the number of sentences in D_i. Given that n_i varies across all languages, we want to ensure that the tokenizer is not overfitted on the languages with more sentences and underfitted on languages with fewer of sentences. Hence, we sample languages according to a multinomial distribution with probabilities dependent on the number of sentences of a language, the total number of sentences in from all languages, and a sample parameter. We follow the distribution introduced in XLM [16] where:

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^{N} p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k}. \tag{3.1}$$

We consider $\alpha = 0.3$ following the results from preliminary results. This ensures that tokens from languages with small number of sentences are well represented in the tokenizer.

Chapter 4

Experimental Setup

4.1 Languages

We focus on 11 African languages, namely Afaan Oromoo (also called Oromo), Amharic, Gahuza (a code-mixed language containing Kinyarwanda and Kirundi), Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya and Yorùbá. These languages all come from three language families: Niger-Congo, Afro Asiatic and English Creole. We select these languages because they are the languages supported by the British Broadcasting Corporation (BBC) News, which was our main source of data. Table 4.1 provides details about the languages used in pretraining our models. As we can see, these languages are collectively have over 400 million speakers.

4.2 Data

4.2.1 Pretraining Data

Working with a research colleague [44], we obtained most of the data from the British Broadcasting Corporation (BBC) News' website.[1] We also obtain some additional data from the Common Crawl Corpus [15, 61] for languages available there, specifically Amharic, Afaan Oromoo, Amharic, Hausa, Igbo, Somali and Swahili. We refer to the corpus as the AfriBERTa corpus. Examples of sentences from each language in the dataset are shown in Figure 4.1.

[1] https://www.bbc.co.uk/ws/languages (scraped up to January 17, 2021)

Language	Family	Speakers	Region	Script
Afaan Oromoo	Afro-Asiatic	50M	East	Latin
Amharic	Afro-Asiatic	26M	East	Geez
Gahuza	Niger-Congo	21M	East	Latin
Hausa	Afro-Asiatic	63M	West	Latin
Igbo	Niger-Congo	27M	West	Latin
Nigerian Pidgin	English Creole	75M	West	Latin
Somali	Afro-Asiatic	19M	East	Latin
Swahili	Niger-Congo	98M	Central/East	Latin
Tigrinya	Afro-Asiatic	7M	East	Ge'ez
Yorùbá	Niger-Congo	42M	West	Latin

Table 4.1: **Language Information:** For each language, its family, number of speakers [20], and regions in Africa spoken.

Language	XLM-R	mBERT	AfriBERTa
Afaan Oromoo	0.10	-	0.05
Amharic	0.80	-	0.21
Hausa	0.30	-	0.15
Somali	0.40	-	0.17
Swahili	1.60	0.04	0.19
Yorùbá	-	0.06	0.03

Table 4.2: **Comparing Sizes Across Models:** Comparison of the dataset sizes (GB) of languages present in XLM-R, mBERT and AfriBERTa. "-" indicates language was not present in model's pretraining corpus.

Size

The total size of the AfriBERTa corpus is 0.94 GB (108.8 million tokens). In comparison, XLM-R was pretrained on about 2395 GB (164.0 billion tokens) [15], and mBERT was trained on roughly 100 GB (12.8 billion tokens).[2] Following findings from RoBERTa [35] and XLM-R [15] that more data is always better for pretrained language modelling, our small corpus makes our task even more challenging, and one can already see that our model is at a disadvantage compared to XLM-R and mBERT.

For each language we pretrained on that is present in XLM-R or mBERT, we compare

[2] https://github.com/mayhewsw/multilingual-data-stats/tree/main/wiki

Language	# Sent.	# Tok.	Size (GB)
Afaan Oromoo	410,840	6,870,959	0.051
Amharic	525,024	1,303,086	0.213
Gahuza	131,952	3,669,538	0.026
Hausa	1,282,996	27,889,299	0.150
Igbo	337,081	6,853,500	0.042
Nigerian Pidgin	161,842	8,709,498	0.048
Somali	995,043	27,332,348	0.170
Swahili	1,442,911	30,053,834	0.185
Tigrinya	12,075	280,397	0.027
Yorùbá	149,147	4,385,797	0.027
Total	5,448,911	108,800,600	0.939

Table 4.3: **Dataset Size:** Size of each language in the dataset covering numbers of sentences, tokens and uncompressed disk size.

the size of that language in our dataset to its size in the pretraining corpora of mBERT and XLM-R. From the comparison details in Table 4.2, we can see that XLM-R always has more data for languages present in our pretraining corpus and theirs. In fact, on average, we can see that the size of the language is always at least two times more in XLM-R. For mBERT, we can see that AfriBERTa has more data for Hausa and Yorùbá, which are present in both corpora. However, one would expect that, given that both languages are in the Latin script, there should be enough high-resource transfer to help them outperform our model.

Our corpus contains approximately 5.45 million sentences and 108.8 million tokens. Table 4.3 presents more details about the dataset size for each language. It can be observed that languages like Swahili, Hausa and Somali have the most amount of data, while languages like Tigrinya have very little data, with just about 12,000 sentences.

Preprocessing

We remove lines that are empty or only contain punctuation. Given that there is significant overlap between the African language corpora in Common Crawl and the BBC News data that we crawled, we perform extensive deduplication for each language by removing exact matched sentences. We also enforce a minimum length restriction by only retaining sentences with more than 5 tokens. We observe that the quality of the dataset from Common Crawl is very low, confirming recent findings from Kreutzer et al [30]. Hence,

Language	Sentence
Afaan Oromoo	Teewoodroos uummata Oromoorratti nama garajabbuummaan duguuggaa raawwataa turedha.
Amharic	የአካባቢው ፀጥታ ባለሥልጣናት የግድያው ምክንያት ምን እንደሆነ እንደማያውቁ ተናግረዋል።
Gahuza	Abakorana nawe bakeka ko yishwe kubera akazi kiwe.
Hausa	Za a yi zaben gwamnan jihar Ekiti ne a ranar 14 Ga Yuli.
Igbo	Elu na-agba gburugburu dripper mmiri anwụrụ mmepụta akara.
Nigerian Pidgin	For 2014, she become di first female mayor for di kontri capital, Libreville.
Somali	Wali xaalku waa sidii, ma jirto cid is waydiinaysa sababta wali laysu dilayo.
Swahili	Mafuvu ya vichwa vya waliouawa katika mauaji ya kimbari Rwanda.
Tigrinya	ዝኽሪ ኤርትራውያንን ኢትዮጵያውያንን ኣብ ግረን ፈል ታመር
Yorùbá	Súre fún èmi náà, baba mi.

Figure 4.1: Example of sentences in the pretraining corpus of each language.

we manually clean the data as much as we can by removing texts in the wrong language, while trying to throw out as little data as possible.

Evaluating Pretraining

We take out varying amounts of evaluation sentences from each language's original monolingual dataset, depending on the language's size. Our total evaluation set containing all languages consists of roughly 440,000 sentences. We evaluate the perplexity on this dataset to measure language model performance. However, following XLM-R [15], we continue pretraining even after validation perplexity stops decreasing. Effectively, we pretrain on around 0.94 GB of data and evaluate on around 0.08 GB of data.

4.2.2 Downstream Tasks Evaluation

We evaluate on two tasks: Named Entity Recognition (NER) which is a form of token classification task and News Topic classification which is a form of text classification.

Named Entity Recognition

In this task, we aim to predict the entity class of each token in a sentence. Named entity classes can range from Persons to Locations to Dates. Tokens which are not named entities are labelled as such. We evaluate NER using the MasakhaNER dataset [2]. The dataset covers the following ten languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian Pidgin, Swahili, Wolof and Yorùbá. The authors established strong baselines on the dataset ranging from simpler methods like CNN-BiLSTM-CRF to pretrained language models like mBERT and XLM-R. We use the train, validation and test splits as released by the MasakhaNER [2] authors. Examples of sentences from each language in the dataset are shown in Figure 4.2.

News Classification

In this task, we aim to classify an article into its corresponding news topic. We use the news topic classification dataset from [25], which covers Hausa and Yoruba. The Yoruba dataset has 7 categories, namely "Nigeria", "Africa", "World", "Entertainment", "Health", "Sport", "Politics". The Hausa dataset has 5 categories, which are the same as all the Yoruba dataset categories excluding "Sport" and "Entertainment". The authors established strong transfer learning and distant supervision baselines. They find that both mBERT and XLM-R outperform simpler neural network baselines in few-shot and zero-shot settings. We use the train, validation and test splits as released by the authors [25]. Examples of sentences from each language in the dataset are shown in Table 4.4.

4.3 Implementation

4.3.1 Pretraining

We pretrain on text data containing all languages, sampling batches from different languages. We sample languages such that our model does not see the same language over several consecutive batches. All models are trained with the Huggingface Transformers library [62] (v4.2.1). We also compare variants of AfriBERTa models to each other in a bid to understand how to pretrain multilingual language models in small data regimes. We explore the design space by pretraining variants from the point of view of model architecture. Three factors are taken into consideration: (i) model depth, (ii) number of attention heads and (iii) vocabulary size. We define performance as "good transfer to downstream task".

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፩፰ ዓመት ያሳለፈውን ዝንግን 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	Ẹ́mírà ilú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọdún méjìdínlógún ní orílè-èdè Nàìjíríà

Figure 4.2: Example of named entities in different languages. PER, LOC, and DATE are in colours purple, orange, and green respectively. The original sentence is from BBC Pidgin: https://www.bbc.com/pidgin/tori-51702073. Adapted from Adelani et al, 2021 [2]

Because the NER dataset covers more languages, we select it as the downstream task for comparing these variants. When exploring the design space, we pretrain each model for 60,000 steps and use a maximum sequence length of 512. We pretrain using a batch size of 32 and accumulate the gradients for 4 steps. Optimization is done using AdamW [36] with a learning rate of 1e-4 and 6000 linear warm-up steps. We use float16 operations to speed up training and reduce memory usage. The final models following the design space exploration are pretrained for 460,000 steps with 40,000 linear warm-up steps and then the learning rate is decreased linearly. We pretrain them with a batch size of 32 on 2 Nvidia V100 GPUs and accumulate the gradients for 8 steps.

4.3.2 Downstream Tasks

NER models are trained by adding a linear classification layer to the pretrained transformer model and finetuning all parameters. Following the hyperparameters used in MasakhaNER [2], we train for 50 epochs with a batch size of 16, a learning rate of 5e-5 and also optimize with AdamW. Figure 4.3 illustrates how the NER task is performed using our BERT-based model.

Text classification models are trained by adding a linear classification layer to the

Language	Sentences	Class
Hausa		
	Hukumar Zaben Nigeria Ta Kara Wa'adin Yin Rajista Zuwa Karshen Wata	Politics
	Nijar: An Kammala Taron Tsoffin Shugabannin Afirka	Africa
	Gwamna Rotimi Amaechi Zai Gana Da Bill Gates	Health
	Matsalar Sufuri a Babban Birnin Tarayya Abuja	Nigeria
	Boris Johnson: Ya Zamo Sabon Shugaban Jam'iyyar Mazan Jiya	World
Yorùbá		
	Árwá: Bákan náà ni a kò f Atiku torí dúkìá àjni wà tó f tà	Politics
	Kìnìhún fa èèyàn kan ya ní Nairobi, àdúgbò dàrú	Africa
	Coronavirus: Àìsàn yîí ti ràn dé Amrika, Thialand àti South Korea	Health
	Building Collapse: Ìdí tí ilé fi ń wó nìyí'	Nigeria
	Harry and Meghan: Mí o ààdédé gbé ìgbés láti kúrò nílé ba	World
	Isreal Adesanya fàgbàhàn Kelvin Gastelum ni Atlanta	Sport
	Amojúr tó ń nu bàtà fi wá oúnj òòjù	Entertainment

Table 4.4: Examples of the news topic classification data training sentences in Hausa and Yorùbá

pretrained transformer model and finetuning all parameters. Following random hyperparameter search on the validation data, we train for 25 epochs with a batch size of 32, warm-up steps of 100, learning rate of 5e-5 and optimize with AdamW as well. Figure 4.4 illustrates how the text classification task is performed using our BERT-based model.

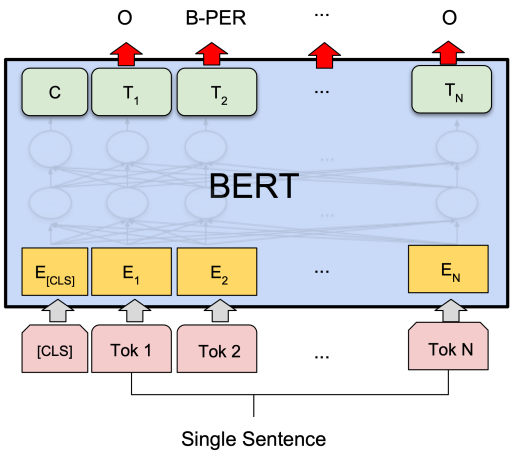

Figure 4.3: Illustration of token classification using BERT: Each token's final hidden state is used for classification for that token. Adapted from Delvin et al, 2019 [18]

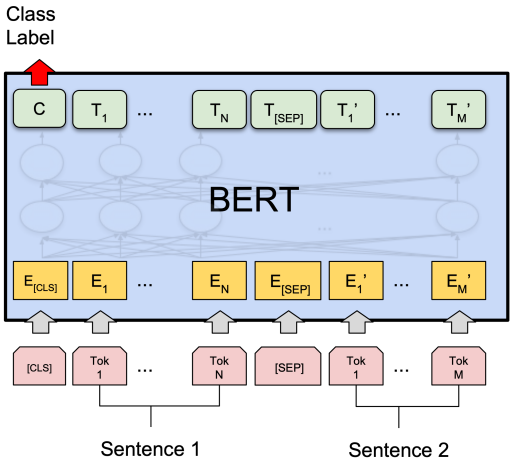

Figure 4.4: Illustration of text classification using BERT: The CLS token (C in diagram) is used for the classification. Adapted from Delvin et al, 2019 [18]

Chapter 5

Results and Analysis

5.1 Design Space Exploration Results

5.1.1 Model Depth

As is common in literature [59, 35, 18, 15, 50], transformer layers are usually in multiples of 2, so we decide to explore the following layer depths: 4, 6, 8 and 10. For each model, we use 4 attention heads and adjust the size of the hidden units and feed-forward layers so that all models have approximately the same number of parameters. From preliminary experiments, models with more than 10 layers did not yield substantially better performance. This is expected, given the small size of the data. Because of this, coupled with computational constraints, we do not explore settings with more than 10 layers.

As we can see from the results in Table 5.1, deeper models always outperform shallower models. However, performance gains diminish with size. For example, the gain from increasing the model to 6 layers from 4 layers is roughly 1 F1 point. However, the gain from increasing from 6 layers to 10 layers is only ∼0.4. This corroborates the recent *universality overfitting* findings from Kaplan et al., [1], who showed that the performance of transformer language models improves predictably as long as data size and model depth are scaled in tandem, otherwise there is a diminishing return.

In general, our results suggest that deeper models also work well when pretraining multilingual language models on small datasets. This follows previous works on understanding the cross-lingual ability of multilingual language models [29], which have shown that deeper models have better cross-lingual performance. However, gains from increasing depth are relatively minimal because of the size of our corpus.

Layers	Params	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
4	74.8M	62.18	89.66	87.03	69.29	67.23	59.00	83.57	83.89	77.04	67.02	75.97
6	74.7M	61.59	90.34	85.81	72.76	66.39	61.43	86.27	84.02	76.61	68.54	76.91
8	74.6M	62.04	90.96	86.33	74.00	68.66	60.96	84.43	84.16	76.11	67.38	77.00
10	74.3M	62.14	90.69	87.36	75.74	67.87	60.59	84.79	84.70	76.17	67.51	**77.27**

Table 5.1: **Effect of Number of Layers:** NER dev F1 scores (averaged over three different random seeds) on each language for models with different layer depth, but same number of parameters. The sizes of the embedding and feed-foward layers are adjusted such that feed-foward is always approximately 4 times embedding size. The highest F1-score per language is underlined, while the highest overall average is in **bold**.

5.1.2 Number of Attention Heads

Again, as is common in literature [59, 35, 18, 15, 29, 39], attention heads are usually in multiples of 2. Hence, for each layer size (4, 6, 8 and 10), we train models with three different numbers of attention heads: 2, 4 and 6. Again, initial experiments with more than 6 attention heads did not yield any better results, so we do not explore more than 6 heads. Results are presented in Table 5.2.

The results suggest that there is a diminishing return to the number of attention heads when the model is deep. Shallower models need more attention heads to attain competitive performance. However, when the model is deep enough, it is very competitive with as few as two attention heads. This suggests that results from recent work [29, 39], which suggest that transformers do not need a large number of attention heads, also hold true for multilingual language models on small datasets.

5.1.3 Vocabulary Size

Previous work has suggested that on small datasets, one should employ a small vocabulary size [56, 7]. However, it remains to be seen if this holds in the multilingual setting since several languages will be competing for vocabulary space and XLM-R [15] have found that increasing the vocabulary size improves multilingual performance. We evaluate our best model size on increasing vocabulary sizes and report results in Table 5.3. As we can see from the results, increasing the vocabulary size does not always yield good results on smaller datasets. While a small vocabulary size performs relatively poorly, medium sized

vocabularies can sometimes outperform larger ones. Due to computation constraints, we selected vocabulary size of 70k for the final models below.

5.1.4 Final Model Selection

We release three AfriBERTa pretrained model sizes: small (4 layers), base (8 layers) and large (10 layers). Each model has 6 attention heads, 768 hidden units, 3072 feed-forward size and a maximum length of 512. Their respective parameter sizes are 97 million, 111 million and 126 million.

5.2 Downstream Task Results

5.2.1 NER Results

As we can see in Table 5.4, even the AfriBERTa small model, which is almost three times smaller than XLM-R, obtains competitive NER results across all languages, trailing XLM-R by less than 3 F1 points. This represents a great opportunity for deployment in resource constrained scenarios, which is usually common for applications in low-resource languages. Our best performing model is AfriBERTa large, which outperforms mBERT and is very competitive with XLM-R across all languages. AfriBERTa large even outperforms both models on several languages that all three models were pretrained on, such as Hausa, Amharic and Swahili.

It should be noted that AfriBERTa large achieves all this with less than half of the number of parameters of XLM-R and about 45M fewer parameters than mBERT. Furthermore, Table 5.5 shows the presence of our test languages in the pretraining corpora of the various models. Our models perform very well on languages that were not part of our pretraining corpus, such as Luo, Wolof and Luganda. This demonstrates its strong cross-lingual capabilities, despite smaller parameter sizes and pretraining corpus size. A notable observation is that both mBERT and XLM-R outperform AfriBERTa on Nigerian Pidgin, despite not being trained on the language. This is likely because of the language's high similarity with English. Nigerian Pidgin is an English Creole, meaning it borrows and shares a lot of its properties (including words) with English. Since both mBERT and XLM-R were pretrained on very large amounts of English data, it is no surprise that they perform so well on Nigerian Pidgin. In summary, our small, base and large models' performance are comparable to mBERT and XLM-R across all languages, despite being pretrained on a substantially smaller corpus and having fewer model parameters.

Layers	Heads	Params	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
4	2	60.1M	58.23	88.78	84.63	71.28	65.68	56.91	83.84	82.44	76.69	64.64	74.99
4	4	60.1M	60.09	89.34	87.08	72.95	68.25	60.10	84.08	83.17	76.29	66.73	76.44
4	6	60.1M	60.26	89.49	86.01	72.69	67.82	59.85	84.68	83.73	76.22	67.66	<u>76.46</u>
6	2	74.3M	60.54	89.72	87.25	72.68	70.23	59.98	84.52	83.25	76.00	67.00	76.74
6	4	74.3M	63.29	90.19	86.05	74.26	68.58	59.23	84.74	83.46	77.62	67.04	76.80
6	6	74.3M	60.38	90.86	86.70	73.12	68.54	61.68	84.59	82.80	79.02	68.48	<u>77.31</u>
8	2	88.5M	60.32	90.55	85.32	75.38	69.89	62.73	85.50	83.51	79.07	68.09	77.78
8	4	88.5M	61.90	90.79	86.67	74.28	68.45	61.57	85.64	83.88	78.48	70.16	77.77
8	6	88.5M	60.92	90.16	86.95	74.71	70.66	60.75	85.48	84.87	78.04	71.16	**<u>78.09</u>**
10	2	102.6M	59.87	90.78	87.10	73.73	66.29	60.03	85.04	83.47	81.12	69.06	77.40
10	4	102.6M	63.95	91.33	87.11	75.24	68.96	63.36	85.66	84.67	74.60	69.27	77.80
10	6	102.6M	63.94	90.54	87.39	75.90	69.19	61.73	85.77	84.66	75.64	69.48	<u>77.81</u>

Table 5.2: **Effect of Number of Attention Heads:** NER dev F1 scores (averaged over three different random seeds) on each language for different models with the same number of layers, but different number of attention heads. The highest F1-score per layer size is <u>underlined</u>, while the highest overall average is in **bold**.

Layers	Heads	Vocab	Params	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
8	6	25k	76.9M	60.56	89.96	85.84	73.23	69.67	61.86	85.11	84.34	75.40	68.35	77.09
8	6	40k	88.5M	60.92	90.16	86.95	74.71	70.66	60.75	85.48	84.87	78.04	71.16	78.09
8	6	55k	99.9M	63.65	90.17	87.28	72.47	67.47	61.49	85.59	85.09	77.56	69.06	77.35
8	6	70k	111.5M	66.17	91.25	87.74	77.44	68.29	59.91	87.00	87.05	77.49	68.82	**78.33**
8	6	85k	123.1M	62.35	90.42	87.44	77.01	68.20	61.98	86.46	85.87	72.84	70.14	77.82

Table 5.3: **Effect of Vocabulary Size:** NER dev F1 scores (averaged over three different random seeds) on the best model size with varying vocabulary sizes. The highest overall average F1-score is in **bold**.

Language	CNN-BiLSTM CRF	mBERT (172M)	XLM-R base (270M)	AfriBERTa small (97M)	AfriBERTa base (111M)	AfriBERTa large (126M)
amh	52.89	0.0	70.96	67.90	71.80	**73.82**
hau	83.70	87.34	89.44	89.01	90.10	**90.17**
ibo	78.48	85.11	84.51	86.63	86.70	**87.38**
kin	64.61	70.98	**73.93**	69.91	73.22	73.78
lug	74.31	80.56	**80.71**	76.44	79.30	78.85
luo	66.42	72.65	**75.14**	67.31	70.63	70.23
pcm	66.43	**87.78**	87.39	82.92	84.87	85.70
swa	79.26	86.37	87.55	85.68	**88.00**	87.96
wol	60.43	**66.10**	64.38	60.10	61.82	61.81
yor	67.07	78.64	77.58	76.08	79.36	**81.32**
avg	69.36	71.55	**79.16**	76.20	78.60	79.10
avg (excl. amh)	71.19	79.50	**80.07**	77.12	79.36	79.69

Table 5.4: **Comparison of NER Results:** F1-scores on the test sets of each language. XLM-R and mBERT results obtained from Adelani et al. [2]. The best score for each language and overall best scores are in **bold.** We also report the model parameter size in parentheses.

5.2.2 Text Classification

We also compare our best model (AfriBERTa large) to XLM-R base and mBERT on text classification. As we can see from the results in Table 5.6, AfriBERTa large clearly outperforms both XLM-R and mBERT by over 10 F1 points on Yorùbá and up to 7 F1 points on Hausa. Results show that mBERT slightly outperforms XLM-R on Yorùbá, most likely because it was pretrained on it, while XLM-R was not. XLM-R also outperforms mBERT on Hausa, presumably for the same reason. It should be noted that our model was pretrained on around half as much Hausa data as XLM-R, but still outperforms it substantially.

5.3 Discussion

In this section, we discuss some other contributions of this work and the implications of the results observed in the previous section. At a high level, AfriBERTa presents the first evidence that multilingual language models are viable with very little training data. This

Language	In mBERT	In XLM-R?	In AfriBERTa?
amh	no	yes	yes
hau	no	yes	yes
ibo	no	no	yes
kin	no	no	yes
lug	no	no	no
luo	no	no	no
pcm	no	no	yes
swa	yes	yes	yes
wol	no	no	no
yor	yes	no	yes

Table 5.5: **Language Presence in pretraining corpora:** This shows the presence of the downstream task test languages in the pretraining corpora of the various pretrained language models.

Language	In mBERT	In XLM-R?	In AfriBERTa?	mBERT	XLM-R base	AfriBERTa large
hau	no	yes	yes	83.03	85.62	**90.86**
yor	yes	no	yes	71.61	71.07	**83.22**

Table 5.6: **Comparison of Text Classification Results:** F1-scores on the test sets. The best score for each language is in **bold**.

offers numerous benefits for the NLP community, especially for low-resource languages.

5.3.1 Opportunities for Smaller Curated Datasets

Our empirical results suggest that state-of-the-art NLP methods like multilingual language models can be made more accessible for low-resource languages. Caswell et al. [30] recently showed that web-crawled multilingual corpora available for many languages, especially low-resource ones, are usually of very low quality. They found issues such as wrong-language content, erroneous language codes and low-quality sentences. Our work opens the door to competitive multilingual language models on smaller curated datasets for low-resource languages.

Another possible benefit of these smaller curated datasets is that they would tend

Model	# Params	Data Size (GB)	# Tokens
XLM-R base	270M	2395	164.0B
mBERT	172M	100	12.8B
AfriBERTa large	126M	0.94	108.8M

Table 5.7: **Comparing Sizes:** Comparison of datasets and model sizes between XLM-R, mBERT and AfriBERTa.

to contain local content as opposed to foreign content as is in the Wikipedia and other relatively larger datasets of these languages. Models trained on such datasets with local content could potentially be more useful to the speakers of the languages given that they would be trained on data with local context.

5.3.2 Strength of Language Similarity

Our work challenges the commonly held belief in the NLP community that lower-resource languages need higher-resource languages in multilingual language models. Instead, we empirically demonstrate that pretraining on similar low-resource languages in a multilingual setting may sometimes be better than pretraining using high-resource and low-resource languages together. This approach should be considered in future work, especially since there have been recent findings [60] that low-resource languages also experience negative interference in multilingual models.

5.3.3 Potential Ethical Benefits

Recent works have called for more considerations of ethics and related concerns in the development of pretrained language models [11]. These concerns have ranged from environmental and financial [57] to societal bias [33, 10]

We believe our work offers the potential to address some of these concerns, while developing language technology for under-served languages. A comparison of model and data sizes of common multilingual models is presented in Table 5.7. Smaller dataset sizes, like ours, mean that these datasets can more easily be cleaned, filtered, analyzed and *possibly* de-biased in comparison to the humongous data sizes of larger language models. We have also shown that smaller-sized models can outperform larger models, despite using smaller training resources. This represents a potential for reduced environmental impact.

While "low-resource" is commonly used in the NLP community to describe a lack of data resources, recent works [42, 4] have argued that "low-resource" also includes a wide range of societal problems, including computational constraints. Thus, our work embodies the broader spirit of "low-resource", as we develop more efficient models on smaller data sizes for under-served languages.

5.3.4 Improving the Representation of African Languages in Modern NLP tools

As discussed in chapter 2, there is very poor representation of African languages in modern NLP tools. Recently, there have been significant efforts towards closing this gap [6, 43, 42, 5, 22, 8, 19, 2]. Our work follows along this path, as there is a need to build language technologies for the over 1.3 billion people on the continent. Besides showing that multilingual language models are viable on low-resource African languages with small training data, we also introduce the first language models for four of these languages: Kinyarwanda, Kirundi, Nigerian Pidgin and Tigrinya. These are four languages with over 50 million speakers [20] who are active users of digital tools. However, these languages have noticeably deficient support in NLP technologies. Our work represents an important step towards improving this.

Chapter 6

Conclusion and Future Work

In this thesis, we show that it is possible to train viable and competitive multilingual pretrained language models on very little data. This is contrary to popular belief in natural language processing literature. We introduced AfriBERTa, a multilingual language model pretrained on less than one gigabyte of data from 11 African languages and show that this model is competitive with models pretrained on larger datasets and even outperforms them on some languages.

In chapter 3, we detail our proposed approach, which is based on the masked language model pretraining scheme of BERT [18]. We also discuss how we train the tokenizer and detail the critical sampling method that ensures the tokenizer can generalize well to all pretraining languages. We perform an extensive design space exploration and detail its setup in chapter 4 and results in chapter 5. Our results suggest that deeper models also work well when pretraining multilingual language models on small datasets. However, gains from increasing depth are relatively minimal because of the size of our corpus. We also find diminishing returns in the number of attention heads. Our comprehensive experiments also highlight important factors to consider when pretraining multilingual language models on smaller datasets.

We evaluate our trained models on two tasks - Named Entity Recognition and text classification. Our evaluation results in chapter 5 show that our models even sometimes outperform larger models (mBERT [18] and XLM-R [15]) on several languages that all three models were pretrained on. More importantly, our model performs well on languages it was not pretrained on. All of this is achieved using tens of millions fewer parameters and training on at least 100 times less data than both these models. We also discuss some practical benefits of viable language models on smaller datasets. We highlight a possible

strength of co-training similar languages and hypothesize that pretraining on similar low-resource languages in a multilingual setting may sometimes be better than pretraining using high-resource and low-resource languages together. Other possible benefits discussed include potential ethical benefits and an improved representation of languages bereft of data in modern language technology tools.

In future work, we could aim to extend this *small-data* pretraining approach to other modalities such as speech, where many of these languages also have limited resource. Another direction is to expand this model to cover more languages, especially those that are linguistically similar. Furthermore, our models seem to do better on text classification than on NER. It would be interesting to investigate if there are certain tasks where larger multilingual models like mBERT [18] and XlM-R [15] always outperform ours, and vice versa. Finally, we would like to improve the performance of this model by incorporating family-level or script-level syntactic features while making sure not to hurt its multilinguality.

References

[1] Scaling laws for neural language models. *arXiv preprint*, abs/2001.08361, 2020.

[2] David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 10 2021.

[3] Željko Agić and Ivan Vulić. 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 2019. Association for Computational Linguistics.

[4] Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333,

Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[5] Orevaoghene Ahia and Kelechi Ogueji. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. *ArXiv*, abs/2003.12660, 2020.

[6] Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 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, May 2020. European Language Resources Association.

[7] Ali Araabi and Christof Monz. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[8] Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. Contextual text embeddings for twi. *ArXiv*, abs/2103.15963, 2021.

[9] Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5360–5369, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy, August 2019. Association for Computational Linguistics.

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?

In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 06 2017.

[13] Andrew Caines. The geographic diversity of NLP conferences, 2019.

[14] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[15] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[16] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[17] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE Transactions on Audio, Speech and Language Processing*, 2021.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 2019. Association for Computational Linguistics.

[19] Bonaventure F. P. Dossou and Mohammed Sabry. AfriVEC: Word embedding models for African languages. case study of Fon and Nobiin. *ArXiv*, abs/2103.05132, 2021.

[20] David M. Eberhard, Gary F. Simons, and Charles D. Fenning. Ethnologue: Languages of the worlds. (twenty second edition), 2019.

[21] Ignatius Ezeani, Ikechukwu Onyenwe, and Mark Hepple. Transferred embeddings for Igbo similarity, analogy, and diacritic restoration tasks. In *Proceedings of the Third Workshop on Semantic Deep Learning*, pages 30–38, Santa Fe, New Mexico, August 2018. Association for Computational Linguistics.

[22] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond English-Centric Multilingual Machine Translation. *arXiv preprint*, abs/2010.11125, 2020.

[23] J. R. Firth. Applications of general linguistics. *Transactions of the Philological Society*, 56(1):1–14, 1957.

[24] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[25] Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 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, November 2020. Association for Computational Linguistics.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[27] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[28] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.

[29] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: an empirical study. *arXiv preprint*, abs/1912.07840, 2019.

[30] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 01 2022.

[31] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[32] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[33] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics.

[34] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 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, November 2020. Association for Computational Linguistics.

[35] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*, abs/1907.11692, 2019.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[37] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.

[38] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[39] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[40] Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online, November 2020. Association for Computational Linguistics.

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

[42] 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. Participatory research for low-resourced machine translation: A case study

in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November 2020. Association for Computational Linguistics.

[43] Kelechi Ogueji and Orevaoghene Ahia. PidginUNMT: Unsupervised Neural Machine Translation from West African Pidgin to English. *ArXiv*, abs/1912.03444, 2019.

[44] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? No Problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[45] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July 2020. Association for Computational Linguistics.

[46] Jeffrey Pennington, Richard Socher, and Christopher Manning. 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, October 2014. Association for Computational Linguistics.

[47] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 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 2018. Association for Computational Linguistics.

[48] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. Alberto: Italian BERT language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR, 2019.

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learn-

ing with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[51] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[52] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[53] Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online), December 2020. International Committee for Computational Linguistics.

[54] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, 2012.

[55] Rico Sennrich, Barry Haddow, and Alexandra Birch. 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 2016. Association for Computational Linguistics.

[56] Rico Sennrich and Biao Zhang. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July 2019. Association for Computational Linguistics.

[57] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics.

[58] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2014.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[60] Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online, November 2020. Association for Computational Linguistics.

[61] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[62] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[63] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[64] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

[65] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[66] Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online, August 2021. Association for Computational Linguistics.

[67] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international Conference on Computer Vision*, pages 19–27, 2015.