ÌTÀKÚRÒSO: EXPLOITING CROSS-LINGUAL TRANSFERABILITY FOR NATURAL LANGUAGE GENERATION OF DIALOGUES IN LOW-RESOURCE, AFRICAN LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

In this study, we investigate the possibility of cross-lingual transfer from a state-of-the-art (SoTA) deep monolingual model (DialoGPT) to 6 African languages and compare with 2 other baselines (BlenderBot 90M, another SoTA, and a simple seq2seq). The languages are Swahili, Wolof, Hausa, Nigerian Pidgin English, Kinyarwanda & Yorùbá. Natural Language Generation (NLG) of dialogues is known to be a challenging task for many reasons. It becomes more challenging for African languages which are low-resource in terms of data. We translate and train on a small portion of the multi-domain MultiWOZ dataset for the languages. Besides intrinsic evaluation (i.e. perplexity), we conduct human evaluation of single-turn conversations using majority voting and measure inter-annotator agreement (IAA) using Fleiss Kappa and credibility tests. The results show that the hypothesis that deep monolingual models learn some abstractions that generalise across languages holds. We observe human-like conversations in 5 out of the 6 languages. It, however, applies to different degrees in different languages, which is expected. The language with the most transferable properties is the Nigerian Pidgin English, with a human-likeness score of 78.1%, of which 35.5% are unanimous. Its credibility IAA unanimous score is 66.7%. The main contributions of this paper include the representation of under-represented African languages and demonstrating the cross-lingual transferability hypothesis. We also provide the datasets and host the model checkpoints/demos on the HuggingFace hub for public access.
Keywords: Chatbots, African Languages, Conversational Systems, Open Domain, Low-resource

1 INTRODUCTION

The ability to generate coherent messages in natural language is one of the foremost signs of intelligence. Hence, the degree of intelligence of an AI system is reflected in its natural language generation capabilities. Over the years, open-domain conversational systems have evolved (Weizenbaum, 1969; Zhang et al., 2020; Roller et al., 2021; Adiwardana et al., 2020; Adewumi et al., 2019). Advances in deep neural networks, such as the Transformer-based architectures, have brought improvements to the field (Devlin et al., 2018a; Radford et al., 2019; He et al., 2020). These models have demonstrated SoTA performances in Natural Language Understanding (NLU) and Natural Language Generation (NLG) tasks (Wang et al., 2019; Gehrmann et al., 2021).

The advancements notwithstanding, challenges still exist with building conversational systems (Jurafsky & Martin, 2020; Zhang et al., 2020). These challenges include technical (Roller et al., 2021) and ethical challenges (Javed et al., 2021). This is more so that many of the models are originally pretrained on English data, though researchers have recently been producing multilingual versions of some of the models (Devlin et al., 2018b; Conneau & Lample, 2019; Xue et al., 2021). Some of these multilingual models, however, have been shown to have poor performance compared to models trained completely on the target language data (Virtanen et al., 2019; Rönnqvist et al., 2019). The challenges get more daunting for languages that do not have sufficient data to train with, usually called low-resource languages (Nekoto et al., 2020; Adewumi et al., 2020; Adelani et al., 2021). Thus, the multilingual versions of the deep models do not cover many of these languages. For example, 4 of the languages in this work are not available in XLM-R (Conneau et al., 2020) nor mBERT Devlin et al. (2018b), 3 are not in mT5 (Xue et al., 2021), none in mBART (Liu et al., 2020), and 2 are not available on Google Translate. This shows many languages are still under-represented.

In a recent work on cross-lingual transferability, Artetxe et al. (2020) suggest that deep monolingual models learn some abstractions that generalise across languages. We investigate this hypothesis by an empirical study, particularly across 6 diverse, low-resource languages across Africa. Is it possible for a model trained in one language to 'speak' other different languages after finetuning on little data? We show that this is possible. We do not finetune/update the default tokenizer with new tokens or words from the target language. Instead, we leverage the default tokenizers of the selected models.

We present results using perplexity (Brown et al., 1992) and human evaluation. We measure IAA scores using credibility tests and Fleiss Kappa (Fleiss, 1971). We obtain results that apparently validate the earlier stated hypothesis and even obtain better human evaluation results for 2 of the languages than what was shown for Swedish in a similar setup by Adewumi et al. (2021). The language most transferable to appears to be Nigerian Pidgin English. 78.1% of its conversations are judged as human-like, which is more than 20% higher than that obtained by Adewumi et al. (2021). We further contribute the codes[1], datasets and model checkpoints/demos for public use on the HuggingFace hub[2]. The rest of this paper is organised as follows. The 'language of study' section (2) presents brief details of the languages; the methodology section (3) describes the experimental setup, models, data and mode of evaluation; the results & discussion section (4) presents the tables of results and evaluation for all the models, including the error analysis; the conclusion section (6) then follows after the related work section (5), which describes the past related literature.

2 Languages of Study

The following 6 African languages were selected for this work based on their diversity and the availability of contributors. They cover countries in West, East, Central and Southern Africa (Heine et al., 2000) and over 239 million speakers combined. Examples of translated sentences for each language are given in the following subsections. The examples were randomly selected from the training set of the English MultiWOZ dataset.

- I have several options for you; do you prefer African, Asian, or British food?
- i want to book it for 2 people and 2 nights starting from saturday.
- That is all I need to know. Thanks, good bye.

2.1 Swahili

Swahili is a Bantu language. It belongs to the Bantu people in the southern half of Africa (Polomé, 1967). It is an official language of the East African Community (EAC) countries. These include: Uganda, Burundi, Kenya, Tanzania, Rwanda, South Sudan and the Democratic Republic of the Congo (DRC). It is a lingua franca of other areas like Malawi, Mozambique, the southern tip of Somalia, and Zambia (Polomé, 1967). There are more than 50 million speakers of the language. [3] It is also one of the working languages of the African Union. Its example sentences from the English ones are below.

- Nina chaguzi kadhaa kwako; unapendelea chakula cha Kiafrika, Kiasia, au Uingereza?
- nataka kuihifadhi kwa watu 2 na usiku 2 kuanzia Jumamosi.
- Hiyo ndiyo yote ninahitaji kujua. Asante, kwaheri.

[1]

[2]

[3] swahililanguage.stanford.edu

2.2 WOLOF

Wolof is spoken in Senegal, Mauritania and the Gambia. More than 7 million people are believed to speak the language[4]. It belongs to the Senegambian branch of the Niger–Congo language phylum, which is the largest language phylum in the world (Heine et al., 2000). Unlike most other languages of the Niger-Congo phylum, Wolof is not a tonal language. Its example sentences from the English ones are below.

- amna ay tanneef yu bari ngir yaw. ndax bëg ngan lekku niit ñu ñull yi, wa asi wala wa angalteer
- soxla jënd ngir ñaari niit ak ñaari guddi mu tambelee gawu
- dedet li rek la soxla. jerejef. ba benen yoon

2.3 HAUSA

Hausa is a Chadic language spoken by the Hausa people. It is mainly within the northern part of Nigeria and the southern part of Niger. It has significant minorities in Cameroon, Chad, and Benin. It is the most widely spoken language within the Chadic branch of the Afroasiatic phylum (Heine et al., 2000). It has more than 40 million speakers[5]. Its example sentences from the English ones are below.

- Ina da zabubbuka da yawa a gare ku; kun fi son abincin Afirka, Asiya, ko Biritaniya?
- Ina so in yi wa mutane 2 da dare 2 farawa daga ranar Asabar.
- Wannan shine kawai abin da nake bukatar sani. Godiya, bye bye.

2.4 NIGERIAN PIDGIN ENGLISH

Nigerian Pidgin English is a grammatically simplified means of communication among the ethnic groups in Nigeria. Its vocabulary and grammar are limited and often drawn from the English language. It is popular among young people (Ihemere, 2006). About 75 million are estimated to speak the language but the exact number is difficult to estimate since it is not an official language[6]. Its example sentences from the English ones are below.

- I get plenty options for you! you prefer African, Asian, or British food?
- I wan book am for 2 people for 2 night for saturday
- na everything wey i need to know. thank you. good bye

2.5 KINYARWANDA

Kinyarwanda is an official language of Rwanda and a dialect of the Rwanda-Rundi language spoken in Rwanda (Heine et al., 2000). It is one of the four official languages of Rwanda. Over 22 miliion people are estimated to speak the language[7]. Its example sentences from the English ones are below.

- Mfite henshi naguhitiramo hari ibiryo bitetse mu buryo bw' Afrika, Aziya, cyangwa Ubwongereza?
- Ndashaka kubika imyanya ku bantu 2 n'amajoro 2 guhera ku wa Gatandatu.
- Ibyo ni byo nari nkeneye kumenya. Urakoze, murabeho.

[4]worlddata.info/languages/wolof.php
[5]britannica.com/topic/Hausa-language
[6]bbc.com/news/world-africa-38000387
[7]worlddata.info/languages/kinyarwanda.php

2.6 Yorùbá

Yorùbá is predominantly spoken in Southwestern Nigeria by the ethnic Yorùbá people (Heine et al., 2000). It is primarily spoken in a dialectal area spanning Nigeria and Benin with smaller migrated communities in Cote d'Ivoire, Sierra Leone and The Gambia. The number of Yorùbá speakers is more than 45 million[8]. Its example sentences from the English ones are below.

- Mo ní awọ́n àṣàyàn púpọ̀ fún ọ; ṣé o fẹ́ràn óunjẹ Áfríkà, Ásíà, tàbí ìlú Gẹ̀ẹ́sì?
- Mo fé ṣe ìwé fún ènìyàn méjì àti fún alẹ́ méjì tí ó bẹ́rẹ̀ láti ọjọ́ Sátìdeé.
- Ìyẹn ni gbogbo ohun tí mo nílò láti mọ̀. O ṣeun, Ó dàbò.

3 Methodology

We compare 3 models: dialogue generative pre-trained transformer (DialoGPT) (Zhang et al., 2020), BlenderBot 90M (Roller et al., 2021) and a simple seq2seq with attention mechanism, based on the ParlAI platform by Miller et al. (2017). Experiments were conducted using a participatory approach (Nekoto et al., 2020) on Google Colaboratory with free GPUs. Some experiments were on a shared DGX-1 machine with 8 × 32GB Nvidia V100 GPUs. The server runs on Ubuntu 18 and has 80 CPU cores. Each experiment was conducted 3 times and the average perplexity (including standard deviation) was obtained. The finetuning/training process for the BlenderBot and the seq2seq models was for about 20 minutes each. Finetuning DialoGPT on each of the datasets for 3 epochs takes less than 20 minutes. We did not do extensive hyperparameter search due to the constraints of time and resources. The decoding scheme across the models was set as top-k (k=100) and top-p (p=0.7). We recognize that the 3 models do not have exactly the same parameters or configuration and are, therefore, not expected to have the same performance. More details of each model are provided in the following subsections.

3.1 Models

3.1.1 DialoGPT

Zhang et al. (2020) introduced 3 sizes of the DialoGPT: the large, medium and small. It is an English pretrained model for open-domain chatbots based on GPT-2. It was trained on 147M turns of Reddit comments. It uses byte-pair encoding (BPE) tokenizer. The medium model is reputed to have the best performance compared to its large and small versions. In this work, however, we use the small version to minimize the problem of overfitting over small datasets. We utilize the pretrained model from the HuggingFace hub (Wolf et al., 2020). The small model has 117M parameters, 12 layers and uses a vocabulary of 50,257 entries. We use a batch size of 2 during finetuning because of memory constraints and perform ablation studies over the conversation context with values of 7 and 14, noting though that larger context sizes will bring memory challenges (Adiwardana et al., 2020).

3.1.2 BlenderBot 90M

The model is a pretrained transformer model loaded from the ParlAI hub (Miller et al., 2017). It has 8 layers, 16 heads, uses Adam optimizer and byte-level BPE for tokenization. It has 87.5M trainable parameters, a batch size of 6 for finetuning and starts with the learning rate of 1e-5. A variant of English Reddit discussions covering a vast range of topics and totaling 1.5B comments was used to train the model. However, the data consists of group discussions instead of direct two-way conversational data.

3.1.3 Seq2Seq

The seq2seq is an encoder-decoder model that is based on the LSTM architecture (Hochreiter & Schmidhuber, 1997) and uses the attention mechanism (Bahdanau et al., 2015). It was trained from scratch on the datasets in order to compare as a baseline. The model has 805,994 trainable parameters and uses a batch size of 64.

[8] worlddata.info/languages/yoruba.php

3.2 DATA

As a result of the scarcity or non-existent dialogue data for most of the African languages in this work, the authors decided to translate an English dialogue dataset. The poll was between Reddit[9] and MultiWOZ (Budzianowski et al., 2018). Most contributors voted in favour of MultiWOZ, though it is from task-oriented dialogues, instead of Reddit because of the high probability of toxic content (Roller et al., 2021). Indeed, in order to address the challenge of toxic comments in dialogues (Dinan et al., 2019), Solaiman & Dennison (2021) advocated for the approach of carefully curating dataset as a safe approach. They observed that the adjustment of a model's behavior is possible with a small, hand-curated dataset. This approach takes ethical considerations into account (Jurafsky & Martin, 2020; Javed et al., 2021). We follow this approach.

MultiWOZ is a collection of human-human written conversations spanning over multiple domains and topics. Its multiple domain/topic coverage, though limited, makes it ideal for open-domain modeling. Indeed, Budzianowski et al. (2018) experimented with it for neural response generation, showing its usefulness across a range of dialogue tasks. Some of the domains covered are hospital, police, attraction, hotel, restaurant, taxi, train and booking. In our work, we extracted and translated the first 1,000 turns from the training set and the first 250 turns each from the validation and test sets. Contributors were to use either of the two possibilities: human translation or machine translation plus human vetting of all translations. Only the Yorùbá language had a couple of data sources of its own[10]. Thus, only 200 turns from the MultiWOZ training set were added to make up the 1,000 turns. The two Yorùbá sources are a mix of short dialogues in different scenarios such as the market, home and school. It is interesting to note that though the data sizes are small, they are still larger than the COPA benchmark dataset available on the SuperGLUE (Wang et al., 2019).

3.3 EVALUATION

3.3.1 HUMAN EVALUATION

We use the observer evaluation method, where evaluators (or annotators) read transcripts of conversation (Jurafsky & Martin, 2020). Similar to the approach in the original work by Zhang et al. (2020), we ask human evaluators to rate single-turn conversations for human-likeness on a Likert scale of 3 entries. The reason is that lack of long-term contextual information is still an existing problem in conversational systems (Zhang et al., 2020). A copy of each language transcript is given to 3 native speakers per language, as evaluators. A total of 32 single-turn conversations are generated per language and 3 credibility test conversations spread out within the transcript to make up 35. Putting more test conversations would have been desirable but we chose to balance it with the attention span of the annotators, as lengthy transcripts demand more time. A random list was generated and used to select the same 32 prompts for all the languages from each test set. Only one model, which had the best perplexity across languages, was used to generate the conversations: DialoGPT c7 x 1,000 (having context size 7 and 1,000 training turns). The transcripts are available for inspection.

The native speakers had a choice of human-like (H), non-human-like (N) or uncertain (U). These are unbiased respondents who are not connected to the translation of the datasets nor did they take part in the training of the models. Three credible evaluations per language were processed for result computation. Simple majority vote decided the annotation of each single-turn conversation. Out of the total (24) received, 6 were not credible. The credibility test fulfils 2 goals: 1) it helps us check if annotators are qualified or paying attention and 2) it helps us determine the IAA in a simple and intuitive way, especially since the tests are homogeneous to the rest of the conversations. This is what we call the credibility IAA unanimous score. Discredited evaluations are the ones that failed 2 or more out of the 3 credibility test conversations by marking them as anything but H. The 3 credibility conversations are prompts and responses directly from the test set instead of generated responses from the model. A simple instruction for every evaluator at the top of the transcript of conversations is given below.

[9]reddit.com/

[10]YorubaYeMi-textbook.pdf & theyorubablog.com

> Below are 35 different conversations by 2 speakers. Please mark each one as Human-like (H) or Non human-like (N) or Uncertain (U) based on your own understanding of what is human-like.

4 RESULTS & DISCUSSION

Table 1 shows the perplexity results for the best performing model per language. All the values seem relatively low. The Nigerian Pidgin English (46.56) is higher than all the others while Yorùbá (8.76) has the lowest in test set results. It should be noted that it is probably not the best to compare perplexities across languages. The Yorùbá data is largely differently sourced from all the others.

We observe from Table 2 that the single-turn conversations in the transcript of the Nigerian Pidgin English language are judged as human-like 78.1% of the time by majority voting (2/3). 35.5% of them are unanimously (3/3) judged as human-like, which is even higher than both the 3-way split (when each annotator voted for each different category) of 15.6% or non-human-like of 6.3%. This is intuitive, since Pidgin English is closely related to the English language. Meanwhile, the Yorùbá transcript has 0% human-like single-turn conversation, 75% non-human-like conversations, 3.1% uncertain and 21.9% are split 3-way. This makes Yorùbá the least language transferable to, among the set. This may be because of a combination of reasons, including the quality of the dialogue data sources, the language's morphology and written accent peculiarities of the language, among others. Wolof, Hausa, Kinyarwanda and Swahili follow after Nigerian Pidgin English with 65.6%, 31.3%, 28.1% and 28.1% of conversations judged as human-like, respectively.

Model	**Perplexity**	
	Dev (sd)	Test (sd)
Pidgin English	37.95 (0.66)	46.56 (1.13)
Yorùbá*	7.22(0.06)	8.76 (0.08)
Hausa	9.92 (0.05)	12.89 (0.04)
Wolof	14.91 (0.30)	25.85 (0.04)
Swahili	9.63 (0)	9.36 (0.03)
Kinyarwanda	10.85 (0)	14.18 (0.08)

Table 1: Perplexity results (lower is better) across the languages for DialoGPT c7 with 1,000 turns. (*different data sources; sd: standard deviation)

Fleiss Kappa scores are not interpretable using the Kappa 2 annotators on 2 classes guide (Landis & Koch, 1977), as our study uses 3 annotators on 3 classes and the Kappa is lower when the classes are more (Sim & Wright, 2005). Indeed, our study confirms the observation made by Gwet (2014) that the interpretation guide may be more harmful than helpful. Perez Almendros et al. (2020) also report how the Kappa score rose from 41% to 61% when the classes were reduced from 3 to 2. We argue that the credibility IAA approach is a more intuitive and reliable measure, at least, in this case because of the homogeneous test conversations.

Model Language	**Scale (majority voting)**				**F. Kappa**
	H (%)	U (%)	N (%)	3-way (%)	
Pidgin English	78.1	0	6.3	15.6	-0.079
Yorùbá	0	3.1	75	21.9	-0.154
Hausa	31.3	6.3	53.1	9.4	0.228
Wolof	65.6	0	31.3	3.1	0.070
Swahili	28.1	15.6	34.4	21.9	0.067
Kinyarwanda	28.1	25	34.4	12.5	0.091

Table 2: Human evaluation results of 3 annotators on 3 classes. The subjective Kappa example of 2 annotators on 2 classes does not apply here since Kappa is lower when classes are more Sim & Wright (2005)

Figure 1 reveals the relationship between credibility IAA unanimous scores and human-likeness scores for the languages. The least credibility IAA unanimous score for any of the languages is 33.3%. This is the score for Yoruba while Pidgin English, Kinyarwanda, Swahili and Hausa achieve

66.7% and Wolof achieves 100%. Expectedly, the credibility IAA unanimous scores correlate with the human-likeness scores.

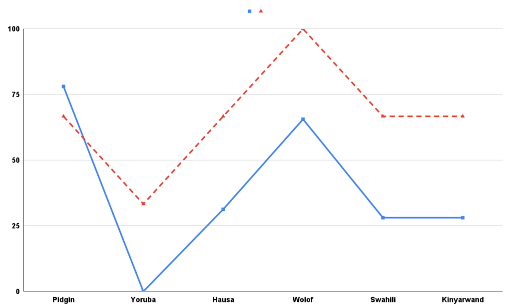

Figure 1: Credibility IAA unanimous scores (red triangle) alongside human-likeness scores (blue square) for the languages

Tables 3, 4, 5, 6, 7 and 8, in the appendix, show the perplexity results across the models for the Nigerian Pidgin English, Yorùbá, Hausa, Wolof, Swahili and Kinyarwanda respectively. It will be observed that for all the languages, the lowest (best) perplexity is achieved with DialoGPT c7 of 1,000 training turns. The only anomaly is with Hausa in Table 5, where the seq2seq model appears to overfit very badly. The results are statistically significant, as all p-values for the difference of two means (between the two lowest results) for all the languages are smaller than the alpha (0.05). Given that these results are obtained with small data, we believe increasing the data sizes will improve the results.

4.1 Error Analysis

Some of the challenges with open domain conversational systems are well-known (Roller et al., 2021; Zhang et al., 2020). We highlight a few of them here. DialoGPT, though the best model in our investigation, shows weaknesses, particularly with the Yorùbá language by failing to generate the correct word morphology, sometimes. Below are 2 examples where the conversations are not coherent.

- *prompt:* Ẹsẹ̀ ń ro mí
- *response:* Eèlóò ni orúkọ ọdún?
- *prompt:* Ìyẹn náà da, wọ́n ṣè ń darúgbó náà ni
- *response:* Kì ni orúkọ́ orúnju rẸ?

On the other hand also, BlenderBot 90M mixes English responses with its generation sometimes, even though the prompts are of the foreign target language. Two examples of prompts in Hausa and responses are given below. There, we also see that it echoes the prompt given to it in the first pair of example.

- *prompt:* Barka dai, me za ku iya gaya mani game da gidan abincin bangkok na burni?
- *response:* barka dai , za za ku iya gaya mani game da gidan abincin bangkok na burni ? i ' m not sure if you ' re being sarcastic or not .
- *prompt:* Ee na mutum 2 da 19:30 na ranar Lahadi don Allah. Ina bukatan lambar nuni kuma.
- *response:* haha , i ' m not sure if i ' m the only one who thinks that .

A major drawback of the seq2seq model is the overly repetitive nature of some of its responses across different prompts. In some case, exactly the same response for two different prompts are generated and the responses are not coherent. We provide feedback from some of the human evaluators in the appendix.

5 Related Work

There are a number of pretrained models for open-domain conversational systems. Some of them include Texar (Hu et al., 2018), DLGnet (Olabiyi & Mueller, 2019), Meena (Adiwardana et al., 2020) and BlenderBot (Roller et al., 2021). These are pretrained on dialogue datasets. In BlenderBot 2 (Komeili et al., 2021; Xu et al., 2021), the same BlendedSkillTalk (BST) (Smith et al., 2020) collection of datasets used for BlenderBot 1 Roller et al. (2021) is used to train the model, in addition to 3 others. There exist, also, models pretrained on large text and adapted for conversational systems. Such models include T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). Another pretrained model on conversational data, DialoGPT, was trained on Reddit conversations of 147M exchanges Zhang et al. (2020). In single-turn conversations, it achieved performance close to that of humans in open-domain dialogues. DialoGPT is based on GPT-2 (Radford et al., 2019). It is an autoregressive model, which achieved SoTA results in different NLP tasks (Radford et al., 2019).

Solaiman & Dennison (2021) observed different harmful outputs in GPT-3, the successor of the GPT-2 model. They discovered that a mitigating factor is carefully curating a small dataset, which determines the behaviour of the model outputs. They made a good case for fine-tuning non-toxic text compared to reducing toxicity through controllable methods using filters or control tokens. Topics such as history, science and government were covered in the dataset (Solaiman & Dennison, 2021). The 80 texts in the values-targeted dataset utilized by Solaiman & Dennison (2021) range in length from 40 to 340 words.

Recently, Artetxe et al. (2020) hypothesised that deep monolingual models learn some abstractions that generalise across languages, while working on cross-lingual transferability. This is in contrast to the past hypothesis that attributes the generalization ability of multilingual models to the shared subword vocabulary used across the languages and joint training, as demonstrated for mBERT (Pires et al., 2019). The performance of such multilingual models on low-resource languages and unseen languages are known to be relatively poor (Pfeiffer et al., 2020; Wang et al., 2021).

In evaluating the performance of open-domain chatbots, it has been shown that automatic metrics, like the BLEU score, can be very poor but they are still used in some cases (Lundell Vinkler & Yu, 2020). Conversation turns per session is another metric of interest Zhou et al. (2020). Perplexity is also widely used for intrinsic evaluation of language models and its theoretical minimum, which is its best value, is 1 (Adiwardana et al., 2020). Probably the best evaluation is done by human evaluators (or annotators) but this can be subjective. The judgment of human evaluators is seen as very important, especially since humans are usually the end-users of such systems (Zhang et al., 2020).

6 Conclusion

In this empirical study of the cross-lingual transferability of a monolingual model for 6 African languages, we observe that it is possible to different degrees of success. The English pretrained DialoGPT model resulted in the best perplexity scores across the languages and provided us the reason to generate single-turn dialogues with it for human evaluation. Nigerian Pidgin English appears to have the most transferable properties and has the best human evaluation results. The hypothesis that deep monolingual models learn some abstractions that generalize across languages appears to hold.

Better performance may be achieved if the tokenizers are optimized on the target languages by training from scratch or fine-tuning, as this will allow more native tokens to be represented. It may be worthwhile to construct a transferability index for various languages. This will indicate the amount of benefit that may be harnessed from utilising such properties in different downstream tasks. Having shown that the cross-lingual transferability hypothesis of deep monolingual models appears to hold, at least for English as a pretraining/source language, is it also possible that other/target languages have such capabilities in a reverse training? This may be an interesting query to investigate.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Chester Lignos, Constantine 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. ISSN 2307-387X. doi: 10.1162/tacl_a_00416. URL `https://doi.org/10.1162/tacl_a_00416`.

Tosin Adewumi, Nosheen Abid, Maryam Pahlavan, Rickard Brännvall, Sana Sabah Sabry, Foteini Liwicki, and Marcus Liwicki. Sm {\aa} prat: Dialogpt for natural language generation of swedish dialogue by transfer learning. *arXiv preprint arXiv:2110.06273*, 2021.

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. Conversational systems in machine learning from the point of view of the philosophy of science—using alime chat and related studies. *Philosophies*, 4(3):41, 2019.

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. The challenge of diacritics in yoruba embeddings. *arXiv preprint arXiv:2011.07605*, 2020.

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL `https://aclanthology.org/2020.acl-main.421`.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations, ICLR 2015*, 2015. URL `https://arxiv.org/pdf/1409.0473.pdf`.

Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Jennifer C Lai, and Robert L Mercer. An estimate of an upper bound for the entropy of english. *Computational Linguistics*, 18(1): 31–40, 1992.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL `https://aclanthology.org/D18-1547`.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf`.

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*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018a.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Multilingual bert. 2018b. URL https://github.com/google-research/bert/blob/a9ba4b8d7704c1ae18d1b28c56c0430d41407eb1/multilingual.md.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *arXiv preprint arXiv:1908.06083*, 2019.

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

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*, 2021.

Kilem L Gwet. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

Bernd Heine, Derek Nurse, et al. *African languages: An introduction*. Cambridge University Press, 2000.

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

Zhiting Hu, Haoran Shi, Bowen Tan, Wentao Wang, Zichao Yang, Tiancheng Zhao, Junxian He, Lianhui Qin, Di Wang, Xuezhe Ma, et al. Texar: A modularized, versatile, and extensible toolkit for text generation. *arXiv preprint arXiv:1809.00794*, 2018.

Kelechukwu Uchechukwu Ihemere. A basic description and analytic treatment of noun clauses in nigerian pidgin. *Nordic journal of African studies*, 15(3):296–313, 2006.

Saleha Javed, Tosin P Adewumi, Foteini Simistira Liwicki, and Marcus Liwicki. Understanding the role of objectivity in machine learning and research evaluation. *Philosophies*, 6(1):22, 2021.

D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Dorling Kindersley Pvt, Limited, 2020. ISBN 9789332518414. URL https://books.google.se/books?id=ZalcjwEACAAJ.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. *arXiv preprint arXiv:2107.07566*, 2021.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pp. 159–174, 1977.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

Mikael Lundell Vinkler and Peilin Yu. Conversational chatbots with memory-based question and answer generation, 2020.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 79–84, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-2014. URL https://aclanthology.org/D17-2014.

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*, pp. 2144–2160, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL https://aclanthology.org/2020.findings-emnlp.195.

Oluwatobi Olabiyi and Erik T Mueller. Multiturn dialogue response generation with autoregressive transformer models. *arXiv preprint arXiv:1908.01841*, 2019.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5891–5902, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.518. URL https://aclanthology.org/2020.coling-main.518.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pp. 46–54, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.7.

Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493.

Edgar C Polomé. Swahili language handbook. 1967.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

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 learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL https://aclanthology.org/2021.eacl-main.24.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. Is multilingual bert fluent in language generation? *arXiv preprint arXiv:1910.03806*, 2019.

Julius Sim and Chris C Wright. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268, 2005.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2021–2030, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.183. URL https://aclanthology.org/2020.acl-main.183.

Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. 2021. URL https://proceedings.neurips.cc/paper/2021/file/2e855f9489df0712b4bd8ea9e2848c5a-Paper.pdf.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*, 2019.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*, 2021.

J Weizenbaum. A computer program for the study of natural language. *Fonte: Stanford: http://web. stanford. edu/class/linguist238/p36*, 1969.

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*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*, 2021.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 270–278, 2020.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.

A Appendix

Model	Training Turns	Perplexity	
		Dev (sd)	Test (sd)
DialoGPT c7	500	42.55 (0)	52.81 (0)
DialoGPT c14	500	67.57 (2.53)	90.18 (3.24)
DialoGPT c7	1,000	**37.95** (0.66)	**46.56** (1.13)
DialoGPT c14	1,000	70.21 (2.17)	92.23 (2.33)
BlenderBot	1,000	81.23 (0)	81.23 (0)
Seq2seq	1,000	277.2 (15)	277.2 (15)

Table 3: Nigerian Pidgin English results (c7 & c14: context sizes 7 & 14; sd: standard deviation)

Model	Training Turns	Perplexity	
		Dev (sd)	Test (sd)
DialoGPT c7	500	10.52 (0.04)	9.65 (0.01)
DialoGPT c14	500	12.63 (0.47)	10.66 (0.4)
DialoGPT c7	1,000	**7.22** (0.06)	**8.76** (0.08)
DialoGPT c14	1,000	7.63 (0.13)	9.11 (0.14)
BlenderBot	1,000	154.43 (0.06)	154.43 (0.06)
Seq2seq	1,000	45.85 (1.41)	45.85 (1.41)

Table 4: Yorùbá results (c7 & c14: context sizes 7 & 14; sd: standard deviation)

Model	Training Turns	Perplexity	
		Dev (sd)	Test (sd)
DialoGPT c7	500	18.53 (0.23)	25.7 (0.4)
DialoGPT c14	500	26.40 (0.75)	35.95 (0.73)
DialoGPT c7	1,000	**9.92** (0.05)	**12.89** (0.04)
DialoGPT c14	1,000	11.30 (0.04)	15.16 (0.05)
BlenderBot	1,000	39.39 (1.61)	39.39 (1.61)
Seq2seq	1,000	1.92 (0.12)	1.92 (0.12)

Table 5: Hausa results (c7 & c14: context sizes 7 & 14; sd: standard deviation. Its seq2seq appears to overfit badly)

A.1 Post-Human-Evaluation Feedback

Some of the evaluators, in a post-evaluation feedback, explained that coherence of the conversation mattered as a deciding factor in their judgment. So did the grammar. For example, considering Yorùbá, responses that referenced inanimate objects as if in the context of animate objects or humans were voted as non-human-like. For Wolof, many of the conversations are human-like except for cases where the responses were inconsistent with the prompt or question given. For example, there were conversations that were hard to judge because the responses are questions to the prompts, which happen to be questions themselves. Such conversations were awarded the uncertain votes by the particular annotator.

Model	**Training Turns**	**Perplexity**	
		Dev (sd)	Test (sd)
DialoGPT c7	500	15.2 (0.09)	26.41 (0.10)
DialoGPT c14	500	15.2 (0.09)	26.41 (0.10)
DialoGPT c7	1,000	**14.91** (0.3)	**25.85** (0.04)
DialoGPT c14	1,000	16.61 (0.2)	30.37 (0.08)
BlenderBot	1,000	108.7 (0)	108.7 (0)
Seq2seq	1,000	401.6 (10.39)	401.6 (10.39)

Table 6: Wolof results (c7 & c14: context sizes 7 & 14; sd: standard deviation)

Model	**Training Turns**	**Perplexity**	
		Dev (sd)	Test (sd)
DialoGPT c7	500	15.55 (0.17)	14.22 (0.14)
DialoGPT c14	500	20.03 (0.29)	17.02 (0.22)
DialoGPT c7	1,000	**9.63** (0)	**9.36** (0.03)
DialoGPT c14	1,000	11.07 (0.04)	10.71 (0.05)
BlenderBot	1,000	128.8 (0.1)	128.8 (0.1)
Seq2seq	1,000	134.5 (2.75)	134.5 (2.75)

Table 7: Swahili results (c7 & c14: context sizes 7 & 14; sd: standard deviation)

Model	**Training Turns**	**Perplexity**	
		Dev (sd)	Test (sd)
DialoGPT c7	500	19.28 (0.19)	21.62 (0.22)
DialoGPT c14	500	24.47 (0.17)	26.45 (0.17)
DialoGPT c7	1,000	**10.85** (0)	**14.18** (0.08)
DialoGPT c14	1,000	12.84 (0.1)	17.43 (0.14)
BlenderBot	1,000	177.87 (0.06)	177.87 (0.06)
Seq2seq	1,000	195.07 (7.66)	195.07 (7.66)

Table 8: Kinyarwanda results (c7 & c14: context sizes 7 & 14; sd: standard deviation)