|  |
| --- |
| **Transfer of English Language Models to Creoles and Other English Relatives** |
|  |
| **Diana Chu, Marc Loeb, Breiten Sundra, Sushan Zhao** |
| <https://github.com/marcdloeb/Creoles_project> |
|  |
|  |
|  |

Abstract

Pidgins and creoles are spoken by hundreds of millions of people in post-colonial states worldwide. Because of social stigma and other constraints, these languages are historically understudied in the sphere of NLP. Model transfers, a topic of interest in domains across the field, offers a potential means to reduce the gap between high and low resource languages. To investigate the prospects of language transfers, not only within a language but between them, we evaluate the performance of two major English language models (BERT and GPT-2) on six low resource English descended creoles (Jamaican, Belizean, Nigerian, Solomon Island, Hawaiian, and Gullah) and an English relative (Scots). While zero shot performance was poor across the board, we find that after a small amount of fine tuning, English language models perform better on English relatives than on non-relatives (Turkish, Hungarian, Swahili and Basque) after an equivalent amount of fine tuning.

1 Introduction and Background

Recent advances in NLP––particularly the development of increasingly accurate large language models––are confined to just 20 major languages (Magueresse et al., 2020). These novel methods depend on enormous training datasets. The majority of languages, which researchers and academics have yet to build corpuses for, are referred to as Low Resource Languages (LRLs).

The NLP resources available for a language are not solely a product of the number of speakers. Many LRLs are primarily spoken by marginalized and minority communities and face varying degrees of social stigma. The underrepresentation of LRLs in NLP and linguistic research has repercussions that extend beyond natural language processing. By ignoring their languages in study, we risk prolonging the marginalization and erasure of these communities that already face significant social and economic challenges. As a result, it is critical for scholars to acknowledge the significance of LRLs and work toward more inclusive and equitable methods in NLP and linguistic research.

Researchers are beginning to take steps in bridging the gap between HRLs and LRLs, finding ways to transfer robust existing models to LRLs. However, this is not a superficial, and still requires some kind of corpus. Even a reduced data requirement may be too high a hurdle to clear for many LRLs.

* 1. Pidgins and Creoles

Pidgins are simplified means of communication that facilitate interaction between disparate linguistic groups. Pidgins are not native languages, but rather are learned as a second language by speakers who already have a first language. Historically, many pidgins have emerged because of trade or conquest. The population transfers and commercial interactions caused by European colonialism gave rise to a vast number of pidgins across Asia, Africa, and the America, which mixed European and indigenous linguistic structures. Due to their origins many viewed pidgins as unsophisticated versions of their original languages. In actuality, each pidgin has its own rules for usage that must be learned to become proficient in it.[[1]](#footnote-2) Pidgins in use for a long period of time can become increasingly complex in its grammar and vocabulary. They can then evolve into creole languages when they become the primary language of a community.

Creoles are nativized pidgin. Spoken by hundreds of millions of people across the post-colonial world, they have a mixed linguistic heritage, incorporating elements from multiple source languages culminating into a single creole. Creoles can sometimes be difficult to identify and classify because they often share features with both their source languages and other creole languages. Even though creoles are more developed than pidgins, there are many creole languages that are referred to as pidgins (Nigerian, Solomon Islands, Hawaiian pidgins, etc.). Creoles spoke in West Africa and the Caribbean have a complex morphology comparable to those of HRLs (Lent et al., 2022).

Because of their genesis among subaltern populations of colonial empires, creoles and pidgins have long been stigmatized as improper versions of their parent languages. This is true in much of academia, and in education, where students are penalized or discouraged from using their own language variety, instead being forced to conform to the norms of a standardized HRL (Siegel, 1999). The stigma surrounding these languages has led to a lack of understanding and recognition of the unique linguistic and cultural contributions that these varieties can offer.

* 1. Language Model Transfer

Language model transfer describes various mechanism to leverage data from domains distinct from the NLP task being evaluated. Transfer has received increasing attention in recent years, as NLP data demands have stressed the limits even of high resource languages (Ruder 2019). Though often confined to shuffling data between domains within a language, model transfer can provide a means to overcome the linguistic resource gap.

Magueresse et al. (2020) discuss methods to apply models from High Resource Languages (HRLs) to be functional for LRLs, specifically addressing the issue of the lack of data. Rather than manually tagging corpora, automatic alignment projects either words themselves or full sentences from HRLs to LRLs, saving lots of time and effort while also obtaining a workable corpus for the LRL. In particular, POS tagging is something that is difficult to do for LRLs. Automatic alignment performs better than randomly sticking tags on words but has its limitations due to the fundamental differences between languages. A different paper by Fang and Cohn (2017) suggests a different approach, taking a bilingual dictionary, monolingual corpora in HRLs and LRLs, and a small annotated corpus of around 1,000 tokens to more accurately tag LRLs.

NLP studies done on pidgins and creoles have made progress on preparing them for further research. A notable study done on West African Pidgin English, a language spoken by over 75 million people, made improvements in the language model transfer process from English models to West African Pidgin (Chang et al., 2021). Although the improvements were unable to process the text to a point where it could be used for a model, the improvements made show promise for further research.

* 1. English Language Relatives

The British Empire was the largest ever to exist, covering a quarter of the world, and containing a quarter of the global population at its peak. In addition to spreading the English language itself, British colonialism gave rise to a large number of pidgins, many of which have evolved into creoles in Africa, the Caribbean and the Asia-Pacific region. In this project, we examine six English language creoles, and a seventh “sibling” language.

Jamaican Creole/Patois

Jamaican Creole, also known locally as Patois or Patwah, is an English-based creole spoken by around 3.2 million people in Jamaica. It originally developed by contact between English speakers and enslaved West Africans, notably the Akan people. As a result, there is a significant Akan substrate effect, although the primary corpus of the language is mostly English.

Nigerian Pidgin English

Nigerian Pidgin, known locally as Naija, is another English-based creole, spoken in Nigeria as an unofficial lingua franca. It is a special case, in that it does not refer to a single language, but rather something of a dialect continuum. The fact that it serves as a lingua franca within Nigeria means that individual language communities have unique takes in the language and incorporate words from their native language. The version we used for our testing is a kind of consensus variant, meant to be accessible to most groups of speakers.

Gullah

Gullah is a creole spoken primarily in coastal South Carolina in the United States. Despite an increasingly dwindling group of speakers (in the hundreds), it has attracted more academic interest and is betterd resourced than other, more widely spoken creoles in this project.  Gullah possesses a mixed substrate from both West and Central African languages, notably from the Bantu family.

Hawaiian Pidgin English

Despite the name Hawaiian Pidgin is a fully developed creole language spoken in Hawaii in the United States. It originated as a language spoken on sugarcane plantations as a means of communication between the English-speaking colonists and Hawaiian-speaking plantation workers. Like many creoles, it is spoken on a continuum between heavily-creolized and more standard English. The version we are using strikes a balance between the two.

Belizean Creole

Belizean Creole is a language spoken by the people of Belize, and developed under influence from English as well as several other creoles and languages, including Jamaican patois. It is also significant in that it includes a large substrate effect from Spanish as well, as many of Belize’s neighbors are Spanish-speaking. It also incorporates many words and expressions from other languages spoken in the region, such as Maya and Garifuna.

Solomon Islands Pijin

Solomon Islands Pijin is an English-based creole language spoken in the Solomon Islands. It is often considered a dialect of New Guinea Tok Pisin, which has more speakers and is slightly better resourced. Unlike the creoles discussed above, Pijin initially emerged in a trade context rather than one of explicit labor or colonization, though this did occur in later stages of development. As a consequence, vocabulary taken in at different stages of development differs internally.

Scots

Unlike the rest of the languages on this list, Scots is neither a pidgin nor a creole, though it is a generally low-resource language. Scots is a language which developed from Middle English and is currently spoken in Scotland. It is important to note that Scots is not the same as Scottish English. The latter is simply English spoken with certain Scottish characteristics and in a Scottish accent. Scots is a full-fledged language like the others on this list, with unique grammar and vocabulary which evolved in a separate direction from standard British or American English. It is one of the few non-creole languages to have a high degree of mutual intelligibility with English.

1. Methods
   1. Data

The fundamental constraint we faced in the course of this project was a lack of data. Low resource languages are, by definition, low resourced. The papers described above are all that we could find in several weeks of intense search. The domains of the handful of datasets that do exist are scattered, ranging from parts of speech to sentiment, which makes performance from language to language comparison difficult.

As a consequence of this scarcity we chose to focus on unlabeled data and unlabeled evaluation metrics. To build a dataset that would maximize our ability to compare performance across languages, we turned to the world’s most widely translated work: The Bible. As of 2022, the entire Bible has been translated in 724 languages. The New Testament into a further 1617. Many of these languages are creoles. Christian missionaries were highly active in the parts of Africa, the Americas and the Asia-Pacific subject to European colonialism, and were quick to translate the Bible into local languages, and newly developing creoles.

To assemble our data, we scraped Bible.com with the Beautiful Soup package in Python (the site provides no direct download). Published by the American evangelical Life Church, in coordination with the global Bible Society, this site possessed the only online, plain text translations of the Bible we could find for many low resource languages. Ultimately, we assembled the New Testament for seven English language relatives: Scots, Gullah, and Jamaican, Nigerian, Solomon Islands, Belizean and Hawaiian Creoles. As a control we did the same for two English language New Testaments (the 18th century edition of the King James Bible, and the modern New Revised Standard Edition), and four from non-Indo-European languages: Basque, Turkish, Hungarian and Zanzibar Swahili. All thirteen were cleaned and split into individual sentences, forming the basis for perplexity calculations and next sentence prediction, two methods to evaluate model performance on unlabeled data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Language | New Test. Sents. | Zero Shot Test Sents. | Fine Tuning Sents. | Fine Tuned Test Sents |
| Belizean | 12021 | 24041 | 9016 | 6011 |
| Gullah | 15023 | 30045 | 11267 | 7512 |
| Hawaiian | 14794 | 29587 | 11096 | 7397 |
| Jamaican | 12350 | 24699 | 9263 | 6175 |
| King James | 7102 | 14203 | 5327 | 3551 |
| NRSV | 7726 | 15451 | 5795 | 3863 |
| Nigerian | 9297 | 18593 | 6973 | 4649 |
| Scots | 6065 | 12129 | 4549 | 3033 |
| Solomon Islands | 13316 | 26631 | 9987 | 6658 |
| Basque | 7030 | 14059 | 5273 | 3515 |
| Hungarian | 7070 | 14139 | 5303 | 3535 |
| Swahili | 6302 | 12603 | 4727 | 3151 |
| Turkish | 9232 | 18463 | 6924 | 4616 |

* 1. GPT-2 and Perplexity

Perplexity was our first evaluation metric for English language model transfer. For this task, we engaged GPT-2 Large, a large-scale, neutral network-based language model developed by OpenAI in 2019. A unidirectional Transformer model, GPT-2 was trained on a massive corpus of unlabeled English text. It uses a self-attention mechanism to model the relationships between words in a sentence, including longer-range dependencies. The Large version of the model is particularly adept at language translation, text summarization, and text generation, because of its ability to capture complex grammatical structures.

Perplexity––a measure of how well a language model can predict the next word in a sequence of text––is a widely used to evaluate model performance in computational linguistics. Specifically, it is the inverse probability of the test set normalized by the number of words in the test set. A lower perplexity score indicates better performance, as it means the model is better at predicting the next word in the sequence. In the context of model transfer, perplexity values will speak to the ability of GPT-2 to interpret English descended creoles based only on its English language training.

* 1. BERT and Zero Shot Next Sentence Prediction Accuracy

BERT, which stands for Bidirectional Encoder Representations from Transformers is a language representation model which takes a large amount of data through bidirectional conditioning. The pre-trained BERT model can then be fine-tuned on a specific task with a smaller amount of labeled data, making it adaptable to a wide range of NLP tasks, including text classification, question answering, and language generation.

As a bidirectional model, BERT cannot be used to calculate perplexity. However, next sentence prediction (tasking the model to evaluate whether a pair of sentences appear sequentially in text) is one of the two methods used to train BERT on unlabeled data (the other being masking). As such zero-shot NSP accuracy (that is, without any fine tuning) is a means to assess how well BERT’s training generalizes to unfamiliar text.

To implement BERT, we again used PyTorch, calling in the BERTForNextSentencePrediction model from the Transformers module. Every sentence in the corpus for each of our languages was used in two testing datasets, one where the second sentence followed the first, the other with second sentences randomly selected from the corpus.

* 1. Fine Tuned BERT

Lamis (2021) is one of the few papers we found that employs model transfer specifically in the context of moving an English language model to a lower resourced relative (specifically Scots). Lamis finds that English language models perform poorly on zero shot tests with Scots, but that fine tunining on a “very small” amount of data in the target language (much less than would be expected to learn a totally unrelated language) dramatically improves performance.

To evaluate the impact of fine tuning on NSP performance, we split our dataset 75%/25% into training and testing data respectively. To ensure consistency, the same randomly selected chapters of the Bible were assigned to fine tuning or testing for all thirteen languages. For the fine tuning chapters, we created a mixed dataset containing both correct and incorrect sentence pairs. Each sentence in the test chapters was placed in two datasets, one with correct pairs and the other with incorrect.

1. Results
   1. Perplexity

Perplexity scores varied significantly across different creoles and English relatives. In this output, the language with the lowest perplexity score is English, with a score of 5.784701, as the GPT-2 large model was trained based on English corpus. Solomon Island Pijin and Gullah (creoles that are very distinct from each other) performed the next best. Concerningly, Zanzibari Swahili, a non-Indo-European language with a much more distant relationship to English, had lower perplexity scores than all other English relatives (including the very closely related Scots). This challenges the notion that English language model performance is proportional to linguistic distance from English, at least in terms of perplexity.

Table

Description automatically generated

Figure 1: Perplexities

GPT-2 performed worst on Jamaican, Nigerian and Belizean creoles. As discussed earlier, these are languages are related to each other, and have significant non-Indo-European influences. They also contain a number of pronunciation differences, that show up in the more phoneticized spelling of the languages. However, they are not further away from English than Swahili. Nor do we believe that they share less with English compared to other creoles. Ultimately, this perplexity performance warrants further investigation.

* 1. NSP

Consistent with the results of (2021), zero short NSP accuracy was poor across the broad. BERT showed a strong bias towards identifying true pairs, making this prediction 100% of the time for all seven English relatives. Surprisingly, BERT performed nearly as poorly on the slightly archiaic English of the King James Bible, despite the fact that the KJV is almost certainly in its original training data. BERT only correctly tagged 18% of the incorrect pairs from the contemporary New Revised Standard Edition.

|  |  |  |
| --- | --- | --- |
| Language | Zero Shot Acc:  Correct Pairs | Zero Shot Acc:  Incorrect Pairs |
| Belizean | 1 | 0 |
| Gullah | 1 | 0 |
| Hawaiian | 1 | 0 |
| Jamaican | 1 | 0 |
| Nigerian | 1 | 0 |
| Sol. Islands | 1 | 0 |
| Scots | 1 | 0 |
| Relatives Avg. | 1 | 0 |
| King James | 1 | 0.02 |
| NRSV | 0.98 | 0.18 |
| English Avg. | 0.99 | 0.1 |
| Basque | 1 | 0 |
| Hungarian | 1 | 0 |
| Swahili | 1 | 0 |
| Turkish | 1 | 0 |
| Non-Rel Avg. | 1 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| Language | Fine Tuned Acc:  Correct Pairs | Fine Tuned Acc: Incorrect Pairs | Avg. Acc.  Change  w/ Tuning |
| Belizean | 0.61 | 0.8 | 0.2 |
| Gullah | 0.64 | 0.84 | 0.24 |
| Hawaiian | 0.69 | 0.82 | 0.26 |
| Jamaican | 0.61 | 0.85 | 0.23 |
| Nigerian | 0.69 | 0.78 | 0.23 |
| Solomon Islands | 0.58 | 0.89 | 0.24 |
| Scots | 0.78 | 0.71 | 0.24 |
| Relatives Avg. | 0.65 | 0.81 | 0.24 |
| King James | 0.79 | 0.8 | 0.29 |
| NRSV | 0.79 | 0.8 | 0.21 |
| English Avg. | 0.79 | 0.8 | 0.25 |
| Basque | 0.58 | 0.77 | 0.17 |
| Hungarian | 0.55 | 0.81 | 0.18 |
| Swahili | 0.43 | 0.9 | 0.17 |
| Turkish | 0.51 | 0.81 | 0.16 |
| Non-Relative Avg. | 0.52 | 0.82 | 0.17 |

Fine tuning produced an average 24 percentage point increase in accuracy among the seven English relatives, and a 25 percentage point increase among the two English Bibles. These are particularly substantial considering the 50% baseline expected from random chance.

Fine tuning resulted in an average accuracy increase of 17 percentage points among the four languages unrelated to English. While a greater improvement than we expected, the scale of improve is significantly lower than that of related languages. Almost the entirety of this differential can be attributed to performance on true pairs. While BERT was able correctly tag 65% of true pairs among English relatives, it was only able to do so for 52% of true pairs among non-relatives, barely better than random chance.

1. Discussion and Conclusion

Our results, particularly in NSP performance after fine tuning, indicate that English language models achieve higher performance when transferred to English relatives, than when transferred to non-relatives. Our results are by no means definitive. An obvious next step would be testing how fine tuning affects perplexity values.

Though it did better on English relatives, BERT performed much better than expected on English non-relatives after fine tuning. This may suggest that our fine tuning dataset (75% of the corpus) was too large. We would like to investigate the impact of reducing its size.

Many English creoles are closely related to each others, forming sub-families. Belizean, Jamaican, Nigerian Creoles and Gullah all developed in the context of the Atlantic triangle trade, and New World chattel slavery. Hawaiian shares features with other Polynesian creoles, Solomon Island Pijin with those of Melanesia. As a consequence, fine tuning on multiple creoles may further improve performance.

Finally, the fact that BERT did extremely poorly in zero shot testing a contemporary English Bible raises questions about our domain. While the Bible provides a consistent yardstick across many creoles, the kind of language it uses is very distinct from what is spoke and written in everyday life. The ability for a model to transfer from one version of the Bible to another may not generalize to the language more broadly.

References

Chang, Ernie, David Ifeoluwa Adelani, Xiaoyu Shen, and Vera Demberg. 2021. “Unsupervised Pidgin Text Generation by Pivoting English Data and Self-Training.” ArXiv:2003.08272 [Cs], April. <https://arxiv.org/abs/2003.08272>.

Fang, Meng, and Trevor Cohn. 2017. “Model Transfer for Tagging Low-Resource Languages Using a Bilingual Dictionary.” ArXiv:1705.00424 [Cs], May. <https://arxiv.org/abs/1705.00424>.

Lent, Heather, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. “What a Creole Wants, What a Creole Needs.” ArXiv:2206.00437 [Cs], June. <https://arxiv.org/abs/2206.00437>.

Magueresse, Alexandre, Vincent Carles, and Evan Heetderks. 2020. “Low-Resource Languages: A Review of Past Work and Future Challenges.” ArXiv:2006.07264 [Cs], June. <https://arxiv.org/abs/2006.07264v1>.

Siegel, Jeff. 1999. “Stigmatized and Standardized Varieties in the Classroom: Interference or Separation?” TESOL Quarterly 33 (4): 701. <https://doi.org/10.2307/3587883>.

“Ethnologue- Languages of the World.” Ethnologue (Free All). Accessed March 7, 2023. <https://www.ethnologue.com/>.

Harm Lameris and Sara Stymne. 2021. Whit’s the Richt Pairt o Speech: PoS tagging for Scots. In Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 39–48, Kiyv, Ukraine. Association for Computational Linguistics. <https://aclanthology.org/2021.vardial-1.5>

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer Learning in Natural Language Processing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-5004>

1. <https://en.wikipedia.org/wiki/Pidgin> [↑](#footnote-ref-2)