



Low-Resource Cross-Lingual Adaptive Training for Nigerian Pidgin

Pin-Jie Lin^{*1,2}, Muhammed Saeed^{*1}, Ernie Chang^{*3}, Merel Scholman^{2,4}

¹Saarland Informatics Campus, Germany

²Language Science and Technology, Saarland University, Germany

³Reality Labs, Meta Inc.

⁴ILS, Utrecht University, the Netherlands

{pinjie, musaeed, m.c.j.scholman}@lst.uni-saarland.de, erniechyc@meta.com,
m.c.j.scholman@coli.uni-saarland.de

Abstract

Developing effective spoken language processing systems for low-resource languages poses several challenges due to the lack of parallel data and limited resources for fine-tuning models. In this work, we target on improving upon both text classification and translation of Nigerian Pidgin (Naija) by collecting a large-scale parallel English-Pidgin corpus and further propose a framework of cross-lingual adaptive training that includes both continual and task adaptive training so as to adapt a base pre-trained model to low-resource languages. Our studies show that English pre-trained language models serve as a stronger prior than multilingual language models on English-Pidgin tasks with up to 2.38 BLEU improvements; and demonstrate that augmenting orthographic data and using task adaptive training with back-translation can have a significant impact on model performance.

Index Terms: spoken language understanding, low-resource machine translation, low-resource language

1. Introduction

Over the past few years, there has been an increasing interest in developing spoken language processing systems for low-resource languages such as the Nigerian Pidgin (Naija) [1]. With a population of 75 million people in Nigeria, Nigerian Pidgin is a low-resource language that lacks sufficient data for spoken language processing tasks. Consequently, models tend to underperform when it comes to critical tasks, such as sentiment analysis [2] and machine translation [3]. Additionally, the orthographic variation of low-resource languages presents a challenge for language processing models, which can be addressed by collecting diverse datasets and performing data augmentation using the target language lexicon [4, 5, 6]. The absence of parallel Pidgin data creates a considerable obstacle to training neural models with a high number of parameters. It also poses difficulties for fine-tuning pre-trained models on the tasks involving Pidgin language with limited resources, as seen in spoken machine translation and text classification [7, 8, 9].

In this paper, we mitigate the issues of data scarcity by collecting and releasing a large-scale parallel English-Pidgin corpus (Section 2). English being the lexifier of Pidgin proves to be a useful high resource language for pivoting Nigerian Pidgin

to other languages [10]. Thus, we use this English-Pidgin parallel dataset to train language models. Prior work proposed that using multilingual models can benefit low-resource language settings [11]. However, fine-tuning existing models [12] for specific tasks can be challenging due to their large number of parameters and sensitivity to parameter values. Thus, to more effectively leverage existing pre-trained models, we introduce a cross-lingual adaptive framework which involves two training procedures consisting of continual adaptive training and task adaptive training with back-translation. Our approach is designed to adapt a base model to a new language, making it more effective for low-resource languages.

To this end, we introduce a cross-lingual adaptation framework for fine-tuning existing models to Nigerian Pidgin [12, 13] (Section 3). Specifically, we perform continual and task adaptation by continually pre-training language models for Naija, and then fine-tuning the models [14, 15] for the downstream tasks. In our analysis, presented in Section 4, we found that the English-based model is superior to the multilingual one, indicating the importance of training on data specific to the target language. Additionally, we found that using task adaptive training provides a significant impact on model performance in the low-data setting. Our results suggest that cross-lingual adaptive training is a promising approach for building effective spoken language systems for low-resource languages².

Our main contributions are as follows:

- We release the first large-scale English-Pidgin dataset³ to our knowledge, which consists of 29.73K sentence pairs.
- Using the collected corpus, we trained a baseline machine translation model, and release a corpus with 5 million synthetic sentence pairs generated using this system. We further improve upon this translation model with task adaptive training [16], and demonstrate a significant BLEU improvement of 2.28 and 1.69 for Pidgin-English and English-Pidgin respectively over the baseline model.
- We show that the English-based pre-trained model (T5) [12] outperforms its multilingual variant (MT5) [17] by 2.38 BLEU in English-to-Pidgin translation, demonstrating the superiority of English models over multilingual one on English-Pidgin and Pidgin-English translations.

2. Corpus Collection

While there have been several efforts to create datasets for Pidgin [18, 19], the language still lacks a sufficiently sized dataset

²<https://github.com/muhammed-saeed/CLaT>.

³We release the English-Pidgin dataset and 5 million synthetic parallel corpus at https://drive.google.com/file/d/1GOi0h5yU9XPZFRDYCa-hjRzf_Nx_uRE1/view

* Equal contribution.

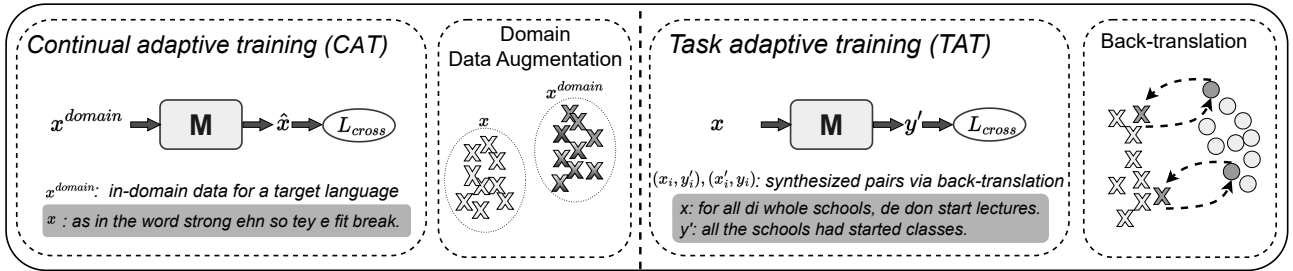


Figure 1: Overview of the framework for low-resource sentiment classification and translation in Pidgin language: (1) **Continual adaptive training**: We consider a base model M and a set of in-domain data x^{domain} in the target language. We then train M with MLM objective which enables a base model to adapt to a new language domain. (2) **Task adaptive training**: Starting from the observed sequence in the source language, the translation model synthesizes an inference in the target language creating the pseudo sentence pair. We construct a bi-directional back-translation by involving the forward and reverse translations. Next, the combined synthetic data serves as a supplementary task for the base model which enables the language model to adapt to more complex tasks via supervised task training.

Table 1: Overview of Pidgin datasets. EN. indicate English language and PG. for Pidgin language. Data included in the corpus, along with their size in a number of sentences.

Corpus	Language	Train	Domain
PARALLEL			
Bible	EN., PG.	29,737	religious
JW300 [18]	EN., PG.	20,218	religious
Naija Treebank [19]	EN., PG.	9,240	misc.
MONOLINGUAL			
NaijaSenti [2]	PG.	8,524	social media
Afri-BERTa [20]	PG.	176,843	news, misc.
BBC Pidgin	PG.	4,147	news
ASR [21]	PG.	7,958	news
PidginUNMT [22]	PG.	5,397	news
IWSLT'15 [23]	EN.	143,609	wiki., misc.
WMT14-En [23]	EN.	4,468,840	news

for application in machine translation models. To address this issue, we combine and enrich various parallel and monolingual texts and datasets to generate a high-quality parallel dataset. The Nigerian Pidgin corpus collection includes six resources: (1) The Holy Bible, where each verse in English was mapped to its corresponding verse in Pidgin, resulting in 29,737 parallel sentences⁴. A limited number of chapters required manual processing to ensure their quality. (2) JW300 corpus, which contains texts from two religious magazines covering various topics. (3) The Naija Treebank, which is a parallel corpus of transcribed spoken Pidgin text with English translations. (4) The NaijaSenti corpus, which consists of 21,017 crawled tweets in Pidgin and three additional Nigerian languages. (5) The Pidgin subset of the Afri-BERTa dataset, which consists of 176K Pidgin sentences, and Pidgin text from 17K Pidgin articles from BBC Pidgin, ASR, and PidginUNMT. (6) 5 million synthetic sentence pairs in English-Pidgin, which were generated from the ISWLT'15 and WMT14 datasets and the Pidgin sentences in the monolingual corpus. Table 1 presents the overview of collected datasets included in the current study along with their respective size.

⁴We utilize the edition provided by Wycliffe Bible Translators, Inc.

Table 2: Types of orthographic variation in Nigerian Pidgin.

Type	Subtype	Example
Alternation	c / k	carry - karry
	a / o	call - coll
Conversion	ou / a	our - awa
	ou / o	your - yor
Transcription	bl / bol	trouble - trobol
	er / a	whether - weda
Deletion	initial	he - e
	medial	different - difren

Orthographic analysis. Due to the lack of a commonly accepted standard orthography in Nigerian Pidgin, we observe various forms of orthographic variation in the data. The data is characterized by both intra-textual variation (i.e. variation within texts from the same source) and inter-textual variation (i.e. between different sources). We identify four main classes of systematic variations that occur in the data: (I) alternation between similar sounds; (II) conversion of digraphs into a single letter or alternate digraphs; (III) phonetic transcription of (blended) letter pairings; and (IV) deletion of silent letters. Table 2 presents examples of each of these classes.⁵

These variations all have phonetic origins. For example, the alternation between “c” and “k” can be attributed to both consonants being ejective, and the conversion of “ee” to “i” can be attributed to both vowels having similar sounds in the Pidgin pronunciation of certain words. As such, we address the inconsistent input by collecting diverse datasets, highlighting the significance of our released data.

3. Cross-Lingual Adaptive Training

Considering the challenges posed by orthographic variations and the scarcity of labeled data for developing performant spoken language processing systems, we introduce two supplementary training approaches—adapting the model to the new lan-

⁵Note that the different datasets adhere to different orthographies – some aim to stay close to English spellings and others aim for phonetic spellings. Both the English spellings and the variations, therefore, occur in our data.

guage and task before fine-tuning on downstream tasks—that can be utilized to benchmark and enhance the performance of low-resource Pidgin sentiment classification and translation tasks: (1) CAT: Continual Adaptive Training and (2) TAT: Task Adaptive Training.

Continual adaptive training. Given the limited availability of labeled Pidgin data, fine-tuning the large number of weights in pre-trained language models (PLMs) is challenging. To this end, we transfer the knowledge about one language absorbed in the weights to the target language by continually adapting the model to a new language via the unlabeled Pidgin corpus. The *Continual Adaptive Training* (CAT) provides supplementary training for the base model to transfer to a specific language domain and thus improves the model’s performance on the downstream task. Figure 1 depicts the training phase where the base model M conducts language adaptation via data assuming from the same domain, thus building an adapted model specialized in a new language. More specifically, an English-based $M^{English}$ is adapted to Pidgin language using large-scale unlabeled data, resulting in a language-specific M^{Pidgin} . Subsequently, we fine-tuned this model for the target tasks.

Task adaptive training. To enhance the model’s ability to tackle more intricate tasks, we further introduce *Task Adaptive Training* (TAT) which allows the model to adapt to the translation task through supervised learning. Our task training involves combining the two sets of synthetic data that possess shared characteristics across both source and target languages for M . To create synthetic data, TAT employs back-translation, a technique that has proven effective in low-resource machine translation scenarios. By leveraging bi-directional back-translation data in our approach, we augment the volume of task-specific training data accessible to the model which can potentially enhance the performance on more complex translation tasks. Specifically, we obtain a synthetic dataset $D'_{x \rightarrow y'} = \{(x, y') | x \in D\}$ via back-translation where the pseudo translation y' was generated according to the sequence x in the source language. We combine two translation directions as the bi-directional back-translation data $D^{BT} = D'_{x \rightarrow y'} \cup D'_{y \rightarrow x'}$.

4. Main Results

General setup. We closely followed the training procedure in transformers [24]. We trained the transformer translation models using Fairseq [25]. For experiments with T5 [12] and MT5 [26], we use Huggingface [27]. We consider BASE for all the checkpoints of the models.

4.1. Sentiment Classification

Data. We derived the low-resource dataset from NaijaSenti [2], which performs sentiment analysis with 3 classes (6.7K/0.6K/1.2K)⁶. We report F1 score.

Setup. We leverage ROBERTA [14] and BERT [15] in base versions. We added INIT baselines where the weights of models are randomly initialized and refer *fine-tuning* as FT which directly transfers the pre-trained language model to Pidgin language. When performing CAT we continually train ROBERTA and BERT on monolingual Pidgin corpus with masked language modeling objective following the instruction in [28], followed by fine-tuning on multi-class classification (“positive”, “negative”, “neutral”) task.

CAT improves Pidgin comprehension. As shown in Table 3, BERT and ROBERTA with continual adaptive training have

⁶We obtained the portion of the dataset from the authors.

both improved FT after the additional pre-training epochs on Pidgin data, resulting in +1 and +2.4 point improvement in F1. Furthermore, CAT enables significant performance gains compared to INIT by +8.9 and +14.1 points of F1. The reason for this can be attributed to poor initialization from INIT where fine-tunes a high number of randomly initialized parameters is challenging, while pre-training and additional adaptive training enable the acquisition of a highly informative language prior to the downstream task.

Table 3: *Results of sentiment classification.*

Model Type	Init	FT	CAT
BERT	71.8	79.7	80.7
RoBERTa	68.4	80.1	82.5

Table 4: *Results on JW300 translation benchmark with data augmentation (DATA AUG.) and task adaptive training (TAT).*

	English-Pidgin	Pidgin-English
<i>Word-level</i>		
JW300 [3]	17.73	24.67
DATA AUG.	23.87	22.61
<i>BPE</i>		
JW300 [3]	24.29	13
DATA AUG.	30.74	28.76
DATA AUG.+TAT	32.43	31.04

4.2. English-Pidgin Translation

Data. We use JW300 translation benchmark [3]. The baseline model uses the JW300 parallel English-Pidgin dataset only. For augmented data, we consider BIBLE which consists of 29K [3]. All models are evaluated on the test set using BLEU score.

Setup. To facilitate a direct comparison with the Pidgin translation benchmark on JW300 [3], we use the identical model architectures for the baselines. The word-level model consists of 4-4 encoder-decoder layers and 10 heads with an embedding size of 300, while BPE model has 6-6 layers, 4 heads, and an embedding size of 256. We performed shared embedding and the shared vocabulary of size 4000. We refer to DATA AUGMENTATION and DATA AUG.+TAT as the model with data augmentation from BIBLE and the model conducting task training on the bi-directional noisy data via back-translation. We exploited back-translation (BT) to produce 430K synthetic parallel sentences from our collected monolingual Pidgin data for TAT. We also release the generated 5 million parallel sentences from the ISWLT15 and WMT14 datasets.

Data augmentation improves performance. Table 4 demonstrates that BPE model with data augmentation significantly improves the baselines by 6.45 and 15.76 BLEU points in both translation directions. For word-level models, augmentation leads to an increase in the BLEU score by 6.14, while the score for Pidgin-to-English translation decreases by 2.06 points. We analyzed the dataset and the model in order to uncover the reason for this decrease, and we found that the *Bible* dataset introduces a lot of orthographic variation when text is segmented at

the word-level while BPE enables sharing more semantic units.

Table 5: Results on JW300 translation benchmark using T5 and MT5.

Model Type	English-Pidgin	Pidgin-English
JW300, BIBLE		
MT5 (BASE)	33.78	32.4
T5 (BASE)	36.16	33.22
ALL		
MT5 (BASE)	33.92	32.75
T5 (BASE)	36.04	34.02
ALL+TAT		
T5 (BASE)	36.35	34.04

TAT with back-translation yields further improvement. As the investigation of TAT’s effectiveness, we generated corresponding parallel sentences by using monolingual Pidgin data with the T5 for 3 epochs of training. Table 4 shows that TAT further improve upon the translation models with the +2.28 and +1.69 BLEU improvement for Pidgin-English and English-Pidgin respectively. This indicates that task adaptive training with back-translation training provides a better initialization for machine translation tasks.

4.3. Further Analysis

English-based model is superior to multilingual models. To validate the hypothesis of the transferability from the English monolingual model and multilingual counterpart for Pidgin language, we compare the T5 where the encoder-decoder is extensively trained on English corpus and the multilingual variant MT5 that was pre-trained on new Common Crawl datasets converting 101 languages. To ensure the fine-tuning of T5 variation models converges smoothly, we train both the base version of T5 and MT5 in the DATA AUG. setting using JW300 and BIBLE. Additionally, we employed ALL the parallel corpus, which consists of BIBLE, JW300, and TREEBANK. Table 5 demonstrates that T5 based solely on the English language outperforms its multilingual counterparts in various scenarios which confirms our hypothesis. We observed a BLEU improvement of +2.38 and +2.12 for both data settings in English-Pidgin translation, while the improvement was +0.82 and +1.27 points in Pidgin-English translation. We concluded that the English-based model is superior to the multilingual one. Moreover, despite using more training data during training, TAT still slightly improves upon T5 baselines. Next, we delve deeper into the potential of task adaptation in improving the adaptability of the base model when faced with limited labeled data.

TAT significantly improves performance in low-data setting. We compare the model with task adaptation stage T5+TAT and the baseline T5 to investigate the impact of task adaptation in low-data scenarios. We used four subsets randomly sampled from the original training splits (20%, 40%, 60%, and 80%) in addition to the full training set. The experimental setting was consistent with that used for English-based T5. Figure 2 shows that T5+TAT substantially outperforms the baselines across 5 sample sizes. We observed that employing TaT obtain particularly strong performance by +3.48 and +2.64 BLEU improvement for Pidgin-English and English-Pidgin respectively when only 20% of the data is available for training. Further, incorporating supervised task training into the model shows a steady

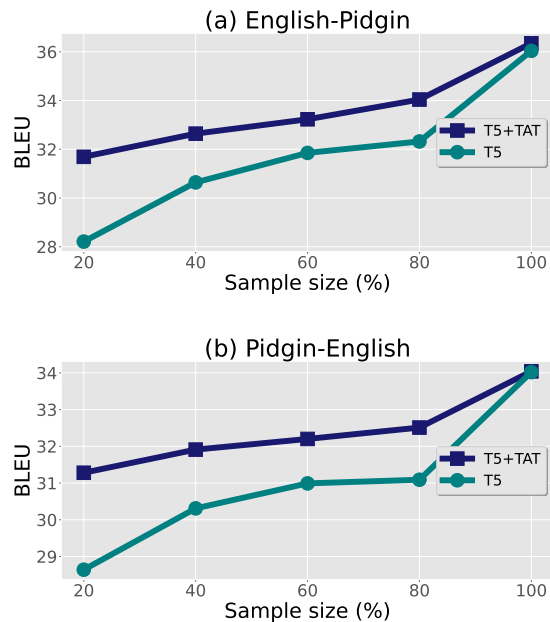


Figure 2: BLEU scores on 20%, 40%, 60%, 80% of sample size and full sample size of (a) English-Pidgin and (b) Pidgin-English translation tasks using T5+TAT framework.

increase across 5 training splits while the performances of the baseline are sensible to the sample size. This indicates the T5+TAT acquired the orthographic information from the task adaptation stage. Thus, the T5+TAT is capable of achieving high performance with less labeled data. The findings suggest that a robust initialization of the language model is essential for performing well in scenarios where data availability is limited, which is often the case in low-resource machine translation applications. Overall, these results highlight the potential value of incorporating TAT into models and suggest avenues for further research into optimizing models for limited data scenarios.

5. Conclusion and Future Works

In this research, we developed an effective spoken language processing framework for Pidgin language text, a low-resource language. We collected the largest parallel English-Pidgin corpus, performed large-scale data augmentation, and proposed a framework for cross-lingual adaptive training. Our studies show that the approach outperforms multilingual models and significantly improves model performance. Our results suggest that cross-lingual adaptive training is a promising approach for spoken language processing systems in low-resource language. For future work, we aim to improve upon the adaptation techniques by better leveraging the English-based PLMs, and making the finetuning process more parameter-efficient for low-resource scenarios.

6. Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft, Funder Id: <http://dx.doi.org/10.13039/501100001659>, Grant Number: SFB1102: Information Density and Linguistic Encoding.

7. References

- [1] H. C. Lent, E. Bugliarello, M. de Lhoneux, C. Qiu, and A. Søgaard, “On language models for creoles,” *CoRR*, vol. abs/2109.06074, 2021. [Online]. Available: <https://arxiv.org/abs/2109.06074>
- [2] S. H. Muhammad, D. I. Adelani, I. S. Ahmad, I. Abdulmumin, B. S. Bello, M. Choudhury, C. C. Emezue, A. Aremu, S. Abdul, and P. Brazdil, “Naijasenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis,” *arXiv preprint arXiv:2201.08277*, 2022.
- [3] O. Ahia and K. Ogueji, “Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin,” *arXiv preprint arXiv:2003.12660*, 2020.
- [4] T. Bergmanis, A. Stefanovics, and M. Pinnis, “Robust neural machine translation: Modeling orthographic and interpunctual variation,” *CoRR*, vol. abs/2009.05460, 2020. [Online]. Available: <https://arxiv.org/abs/2009.05460>
- [5] I. Feldman and R. Coto-Solano, “Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri,” in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 3965–3976. [Online]. Available: <https://aclanthology.org/2020.coling-main.351>
- [6] B. R. Chakravarthi, P. Rani, M. Arcan, and J. P. McCrae, “A survey of orthographic information in machine translation,” *SN computer science*, vol. 2, no. 4, p. 330, 2021.
- [7] E. Nwafor and A. Andy, “A survey of machine translation tasks on Nigerian languages,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6480–6486. [Online]. Available: <https://aclanthology.org/2022.lrec-1.695>
- [8] W. F. Oyewusi, O. Adekanmbi, and O. Akinsande, “Semantic enrichment of nigerian pidgin english for contextual sentiment classification,” *CoRR*, vol. abs/2003.12450, 2020. [Online]. Available: <https://arxiv.org/abs/2003.12450>
- [9] J. O. Alabi, D. I. Adelani, M. Mosbach, and D. Klakow, “Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning,” in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4336–4349. [Online]. Available: <https://aclanthology.org/2022.coling-1.382>
- [10] E. Chang, J. O. Alabi, D. I. Adelani, and V. Demberg, “Few-shot pidgin text adaptation via contrastive fine-tuning,” in *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4286–4291. [Online]. Available: <https://aclanthology.org/2022.coling-1.377>
- [11] A. Mohammadshahi, V. Nikoulina, A. Berard, C. Brun, J. Henderson, and L. Besacier, “Small-100: Introducing shallow multilingual machine translation model for low-resource languages,” *arXiv preprint arXiv:2210.11621*, 2022.
- [12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [13] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, “Multilingual denoising pre-training for neural machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 726–742, 2020. [Online]. Available: <https://aclanthology.org/2020.tacl-1.47>
- [14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.
- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [16] I. Abdulmumin, B. S. Galadanci, and A. Isa, “Enhanced back-translation for low resource neural machine translation using self-training,” in *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers 3*. Springer, 2021, pp. 355–371.
- [17] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. 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*, 2021, pp. 483–498.
- [18] Ž. Agić and I. Vulić, “JW300: A wide-coverage parallel corpus for low-resource languages,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3204–3210. [Online]. Available: <https://aclanthology.org/P19-1310>
- [19] B. Caron, M. Courtin, K. Gerdes, and S. Kahane, “A surface-syntactic UD treebank for Naija,” in *TLT 2019, Treebanks and Linguistic Theories, Syntaxfest*, 2019.
- [20] K. Ogueji, Y. Zhu, and J. 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*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 116–126. [Online]. Available: <https://aclanthology.org/2021.mrl-1.11>
- [21] D. Ajisafe, O. Adegboro, E. Oduntan, and T. Arulogun, “Towards end-to-end training of automatic speech recognition for nigerian pidgin,” *arXiv preprint arXiv:2010.11123*, 2020.
- [22] K. Ogueji and O. Ahia, “Pidginunmt: Unsupervised neural machine translation from west african pidgin to english,” *arXiv preprint arXiv:1912.03444*, 2019.
- [23] M.-T. Luong and C. D. Manning, “Stanford neural machine translation systems for spoken language domains,” in *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, 2015.
- [24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [25] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, “fairseq: A fast, extensible toolkit for sequence modeling,” *arXiv preprint arXiv:1904.01038*, 2019.
- [26] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, “mt5: A massively multilingual pre-trained text-to-text transformer,” *CoRR*, vol. abs/2010.11934, 2020. [Online]. Available: <https://arxiv.org/abs/2010.11934>
- [27] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: <https://aclanthology.org/2020.emnlp-demos.6>
- [28] K. Ogueji, Y. Zhu, and J. 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*, 2021, pp. 116–126.