Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5114–5122
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

IgboBERT Models: Building and Training Transformer Models for the Igbo Language

Chiamaka Chukwuneke[1,2], Ignatius Ezeani[1,2], Paul Rayson[1] and Mahmoud El-Haj[1]
[1]UCREL NLP Group, Lancaster University
[2]Dept. of Computer Sc. Nnamdi Azikiwe University, Anambra State, Nigeria
{c.chukwuneke, i.ezeani, p.rayson, m.el-haj}@lancaster.ac.uk

Abstract

This work presents a standard Igbo named entity recognition (IgboNER) dataset as well as the results from training and fine-tuning state-of-the-art transformer IgboNER models. We discuss the process of our dataset creation - data collection and annotation and quality checking. We also present experimental processes involved in building an IgboBERT language model from scratch as well as fine-tuning it along with other non-Igbo pre-trained models for the downstream IgboNER task. Our results show that, although the IgboNER task benefited hugely from fine-tuning large transformer model, fine-tuning a transformer model built from scratch with comparatively little Igbo text data seems to yield quite decent results for the IgboNER task. This work will contribute immensely to IgboNLP in particular as well as the wider African and low-resource NLP efforts.

Keywords: Igbo, named entity recognition, BERT models, under-resourced, dataset

1. Introduction

The African continent has over 2,000 languages (Eberhard et al., 2020) and information is stored digitally in many of those languages. To digitally interact in those languages, localisation of computer interfaces and tools are very vital and this leads to the need for Natural Language Processing (NLP) research to build these tools. The African continent is underrepresented in the NLP research (Adelani et al., 2021) despite these numerous languages spoken in its 54 countries. Some of the contributing factors to lack of research in these countries include very few available language resources and computers with computing capacity to handle such research. This limits the development and creation of tools and resources for performing a wide variety of NLP tasks such as named entity recognition (NER), machine translation (MT), information retrieval etc.

This paper focuses on building resources - data set and models - for IgboNER i.e. named entity recognition for Igbo, a language mainly spoken in the south eastern part of Nigeria. Named Entity Recognition is a term defined as the task of identifying names of organizations, people, currency, time, percentage expression and geographic locations in text which was introduced at the sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim, 1995). NER is a very important pre-processing step in key NLP tasks such as Question Answering (Mollá et al., 2006), Information Retrieval (Guo et al., 2009), Automatic Text Summarization (Nobata et al., 2002), Machine Translation (Babych and Hartley, 2003) etc. This work is an extension of the on-going efforts (Onyenwe et al., 2014; Ezeani et al., 2016; Ezeani et al., 2020; Onyenwe et al., 2018), and we present the Igbo part of the MasakhaNER dataset to support Igbo NLP in particular but also to contribute to the African NLP efforts. Secondly, we are interested in finding out to what extent a small language model pre-trained from scratch with the target language compares to existing multilingual transformer models like mBERT and XLM-RoBERTa. The rest of the paper is organised as follows: Section 2 discusses the related work; Section 3 describes the raw data for pre-training and the MasakhaNER data set for fine-tuning; Section 4 is about the experiments; Section 5 we discuss some results; Section 6 is on error analysis and in Section 7 we provide conclusions

2. Related Work

2.1. Igbo Language

The African language Igbo (`ibo:ig`) is one of the three major official languages in Nigeria and also an official minority language in Equatorial Guinea and Cameroon. It belongs to the Benue-Congo group of the Niger-Congo family and has over 25 million speakers (Eberhard et al., 2019). It is the native language of the Igbo people, an ethnic group in eastern Nigeria. Igbo is written in Latin script and has over 30 dialects. Igbo has three orthographies: Lepsius, Africa and *"Ọnwụ"* [1]. Historically, there was a lot of controversy over Igbo orthography which led to the standardization by the Ọnwụ Committee in 1961 (Oraka, 1983; Uchechukwu, 2008). The standard Ọnwụ Orthography has 36 graphemes. It consists of eight vowels [*a e o ọ u ụ i ị*] and twenty-eight consonants [*b gb ch d f g gh gw h j k kw kp l m n nw ny ṅ p r s sh t v w y z*] of which 9 are digraphs. Igbo is a tonal language and is written with diacritics. It is an agglutinative language, a single stem can yield many word-forms by addition of affixes that extend its original meaning

[1] https://en.wikipedia.org/wiki/Igbo_language

(Onyenwe and Hepple, 2016). Agreement on a standardized orthography for the Igbo language is difficult, resulting in writing of Igbo texts with a combination of orthographies. Some location and person names are written with African orthography which is the effect of post-colonial government of Nigeria. For example, *"Ọkụzụ"* a town in Anambra State of Nigeria is frequently written with Africa orthography as *"Awkuzu"*. These variants in the way words are written differently could pose a problem for the IgboNER model.

2.2. Low Resource Named Entity Recognition

A lot of research in other languages, mainly English, has been performed on NER since the inception of the task in MUC-6 conference 1996 ((Lample et al., 2016); (Adelani et al., 2021); (Ratinov and Roth, 2009)). One major problem facing NER tasks in low-resourced scenarios is availability of labelled data (Ruder et al., 2019). Manually labelling large corpora is task intensive, time consuming and expensive. With the recent more data-hungry deep learning approach, it has become a bigger challenge to work in this area. Here, we will focus on recent work on NER for low-resourced languages. Adelani et al. (2021) created high quality data sets of less than 4k sentences each for 10 African languages by manual annotation. Transfer learning and gazetteer approaches were applied in building a model that can recognize named entities for 10 African languages. The model was evaluated on multiple state-of-the-art NER models and showed improvement. ANEA, a tool to automatically annotate named entities based on distant supervision to obtain large amount of training data was presented by (Hedderich et al., 2021). ANEA allows users to add their expertise by allowing a tuning step to improve the automatically annotated data. Evaluations on 16 entity types in the following different languages (Spanish, Yoruba, Estonian, West Frisian) showed an improvement on 14 entity types with the F1-score of average.

Tsygankova et al. (2021)'s study proved that using non-speakers annotation is an alternative to cross-lingual methods for building low-resource NER. One of the reasons for its success is the ability of human non-speaker annotators to make inferences over common sense world knowledge unlike an automatic system. Hedderich et al. (2020)'s work on NER and topic classification showed that data sizes affects performance of models. Transfer learning and distant supervision on multilingual transformer models was evaluated on three African languages: Hausa, isiXhosa and Yorùbá, each with different amounts of available resources. This study achieved the same performance as baselines with little data but not for all the cases.

3. Language Resources

3.1. Data Collection

In this work, we used the MasakhaNER dataset created by the Masakhane Community (Masakhane et al., 2021). The data was obtained from BBC Igbo news[2] and is 3,190 sentences containing 61,668 tokens. Additionally, 8,000 Igbo sentences from an ongoing Lacuna project[3] in the Masakhane community was also used. The contents are from Igbo-Radio and Kaoditaa[4].
We also used 383,449 raw monolingual Igbo sentences from the study by (Ezeani et al., 2020). A large section of the data was collected from the Jehovas Witness Igbo[5] and the contents includes the Bible, more contemporary contents (books and magazine e.g. Teta! (Awake!), UloNche! (WatchTower)). Also collected are contents from BBC-Igbo[6], igbo-radio[7] as well as Igbo literary works (Eze Goes To School[8] and Mmadu Ka A Na-Aria by Chuma Okeke).The table 1 shows the statistics of the raw data used in this work.

3.2. Annotation

We used the BIO (Beginning, Inside, Outside) tagging scheme to label the entities. The entity tags correspond to this list: "O", "B-PER", "I-PER", "B-ORG", "I-ORG", "B-LOC", "I-LOC", "B-DATE", "I-DATE" where O denotes non-entity words, B-PER/I-PER denotes the beginning of/is inside a person entity, B-ORG/I-ORG denotes the beginning of/is inside an organization entity, B-LOC/I-LOC denotes the beginning of/is inside a location entity, and B-DATE/I-DATE denotes the beginning of/is inside a date entity. Throughout, 'B' indicates the beginning of a tag, 'I' indicates inside of a tag and 'O' indicates outside i.e. the token belongs to no tag. The annotation of the IgboNER which was performed using the ELISA tool (Lin et al., 2018) by Igbo native speakers from the Masakhane community[9] of which the first author is a member. The ELISA tool was used because it provides an interface for annotators to correct their mistakes, making it easy to achieve a high inter-annotator agreement and also provides an entity level F1 score. Training was given to the annotators to ensure high quality annotation. Fleiss Kappa (Fleiss, 1971) was used to calculate the inter-annotator agreement and it considers each span that an annotator proposed as an entity. The data set has an inter-annotator agreement of 0.995 and 0.9830 at token and entity level respectively. The annotators annotated four entity tags/types: personal name entity (PER), location entity (LOC), organization (ORG), and date &

[2] https://www.bbc.com/igbo

[3] `https://github.com/Chiamakac/lacuna_pos_ner/tree/main/language_corpus/ibo`

[4] https://kaoditaa.com/

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

[6] https://www.bbc.com/igbo

[7] https://igboradio.com/

[8] https://bit.ly/2vdGvKN

[9] https://www.masakhane.io

Source	Sentences	Tokens	Orthography
eze-goes-to-school.txt	1272	25413	Ọnwụ
mmadu-ka-a-na-aria.txt	2023	39731	Ọnwụ
bbc-igbo.txt	34056	566804	Africa, Ọnwụ
igbo-radio.txt	5131	191450	Lepsuis, Africa, Ọnwụ
jw-ot-igbo.txt	32251	712349	Lepsuis, Ọnwụ
jw-nt-igbo.txt	10334	253806	Lepsuis, Ọnwụ
jw-books.txt	142753	1879755	Lepsuis, Ọnwụ
jw-teta.txt	14097	196818	Lepsuis, Ọnwụ
jw-ulo-nche.txt	27760	392412	Lepsuis, Ọnwụ
jw-ulo-nche-naamu.txt	113772	1465663	Lepsuis, Ọnwụ
igbo-radio.txt	2120	11173	Lepsuis, Africa, Ọnwụ
kaoditaa.txt	5880	22557	Lepsuis, Africa, Ọnwụ
Total	391,449	5757931	

Table 1: Data Sources and Counts

time (DATE) using the MUC-6 annotation guide[10]. The annotated entities were based on the state of the art English CoNLL2003 Corpus (Tjong Kim Sang, 2002) but the Miscellaneous (MISC) tag was replaced with the DATE tag in the MasakhaNER data set following previous work (Alabi et al., 2020). The histogram below shows the distribution of the entities annotated.

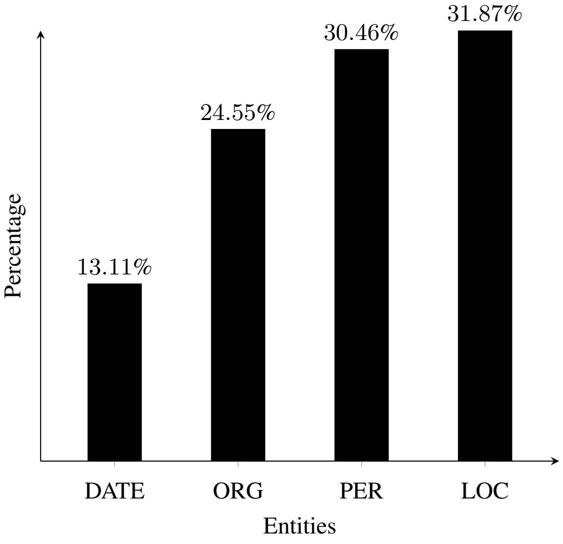

Figure 1: Annotated entity distribution. This shows the percentage distribution of the entities: person (PER), location (LOC), organization (ORG), and date (DATE)

The major issues faced when annotating the Igbo language that we discovered during the process are:

- Orthography: Igbo text corpora are written with combination of Lepsuis, Africa and Ọnwụ orthographies.
- Ambiguity: Some Igbo words are relatively ambiguous. For instance, some person names have other meanings, e.g. "Eze" can be the name of a person (proper noun), a part of the human body for chewing (plural noun) and also it can be a male ruler of an independent state (noun).

3.3. Dataset Splits

The data set is split into three parts named: train, development (dev) and test originally and they correspond to the train, validation and test splits (Masakhane et al., 2021). This was used in fine-tuning all the models in this work. Table 2 shows a summary of the data set splits.

Data set	Number of Sentences	Number of Tokens
Training set	2233	42719
Development set	319	6304
Test set	638	12645

Table 2: Summary of dataset splits

4. Experimental Setup

In this section, we describe the baseline model we pre-trained for Igbo language and some state-of-the-art transformer models we fine-tuned to the downstream Igbo NER task.

4.1. Baseline Model

The first experiment was the training of an Igbo language model (IgboBERT) from scratch using transformers and tokenizers to have a baseline model for Igbo language NER[11]. The model was pre-trained with the raw data described in Table 1 with a masked language modeling (MLM) objective. We trained a byte-level Byte-pair encoding tokenizer (the same as GPT-2) of size 52,000, with the same special tokens as RoBERTa[12]. Our tokenizer is optimized for Igbo by

[10] https://cs.nyu.edu/$\sim grishman$/muc6.html

[11] https://huggingface.co/blog/how-to-train

[12] https://huggingface.co/docs/transformers/model_doc/roberta

encoding native words and diacritics in Igbo language characters. Byte-level Byte-pair encoding tokenizer was chosen because it starts building its vocabulary from an alphabet of single bytes, so all words will be broken down into tokens to eliminate unknown (<unk>) tokens. The small model consists of 6 layers with 768 hidden size, 12 attention heads and 84M parameters, the same number of layers and heads as DistilBERT. IgboBERT was trained at a learning-rate of 1e-4 for 5 epoch and a batch size of 16. We carried out only 5 epoch training because of limited compute resource such as GPU at the time of the experiment. We then fine-tuned the IgboBERT model on IgboNER downstream task using our MasakhaNER dataset.

4.2. Fine-tuned Models

The following state-of-the-art transformer models pre-trained on raw texts only were fine-tuned to a downstream IgboNER task using the MasakhaNER dataset. We added a linear classification layer to the pre-trained transformer models to predict entity types. 20 epoch training with a batch size of 8 at a learning rate of 2e-5 and 1e-4 was run.

- Multilingual BERT (mBERT): mBERT (Devlin et al., 2019) is a transformers model pre-trained with a large corpus of multilingual data from Wikipedia on 104 languages including only two African languages: Swahili and Yorùbá. This model was trained with two objectives: masked language modeling (MLM) and Next sentence prediction (NSP). We use the mBERT-base cased model with 12-layer Transformer blocks consisting of 768-hidden size and 110M parameters

- XLM-RoBERTa (XML-R): XML-R (Conneau et al., 2020) is a multilingual version of RoBERTa pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages including three African languages: Amharic, Hausa, and Swahili. We use of the XLM-R base model consisting of 12 layers, with a hidden size of 768 and 270M parameters.

- DistilBERT (Sanh et al., 2019): a smaller and faster version of BERT, which was pre-trained on the same corpus as BERT. We use the DistilBERT base model uncased with 6-layer Transformer blocks consisting of 768-hidden size and 66M parameters

5. Results

Figure 2 and Figure 3 shows the fine-tuned results of mBERT, XML-R, DistilBERT, and IgboBERT. Table 3 shows the Precision, Recall, F1-score and Accuracy of the fine-tuned models at the learning rate of 1e-4 while Table 4 is the Precision, Recall, F1-score and Accuracy at learning rate of 2e-5. mBERT, XML-R and DistilBERT perform better at the learning rate of 2e-5. IgboBERT at 1e-4 learning rate produced an F1 score 77.94%, Recall 79.50%, Precision 76.44% which showed a better performance over the learning rate of 2e-5 at F1 score of 75.30%, Recall 77.50%, Precision 73.23% but there was no convergence in the training vs. validation loss as seen in Figure 2 which is not good for the model as it shows over-fitting. A solution to this challenge could not be handled in this work and will be explored in further studies. Comparing the results in Table 3 and Table 4, mBERT consistently outperformed at F1-score of 86.66%, 89.02% and accuracy of 97.96%, 98.05% respectively. Our IgboBERT was outperformed by the other models as shown in the tables even though the accuracy level is comparative to others.

6. Error analysis

Figure 4 provides the confusion matrix of the Igbo models which gives a holistic view of the performance of the models. We have the actual values on the x-axis and the predicted values on the y-axis. We can also, see from the matrix that we have more of the non-entity words 'O'.

Model	Precision	Recall	F1	Accuracy
mBERT	85.67	87.67	86.66	97.96
XLM-R	84.54	85.67	85.10	97.81
DistilBERT	79.79	77.00	78.37	96.20
IgboBERT	76.44	79.50	77.94	95.61

Table 3: Performance of mBERT, XML-R, DistilBERT and IgboBERT: We display the fine-tuned results of the models after 20 epoch at 1e-4 learning rate.

Model	Precision	Recall	F1	Accuracy
mBERT	88.22	89.83	89.02	98.05
XLM-R	87.21	88.67	87.93	97.74
DistilBERT	81.26	79.50	80.37	96.67
IgboBERT	73.23	77.50	75.30	95.55

Table 4: Performance of mBERT, XML-R, DistilBERT and IgboBERT: We display the fine-tuned results of the models after 20 epoch at 2e-5 learning rate.

7. Conclusion

We developed a model, IgboBERT, which to the best of our knowledge is the first and only transformer based language model pre-trained on the Igbo Language. We fine-tuned it on a downstream NER task with the masakhaNER data set. Even though the IgboBERT was outperformed as shown by the various F1 scores results in the tables above, we can argue that IgboBERT achieved good performance based on

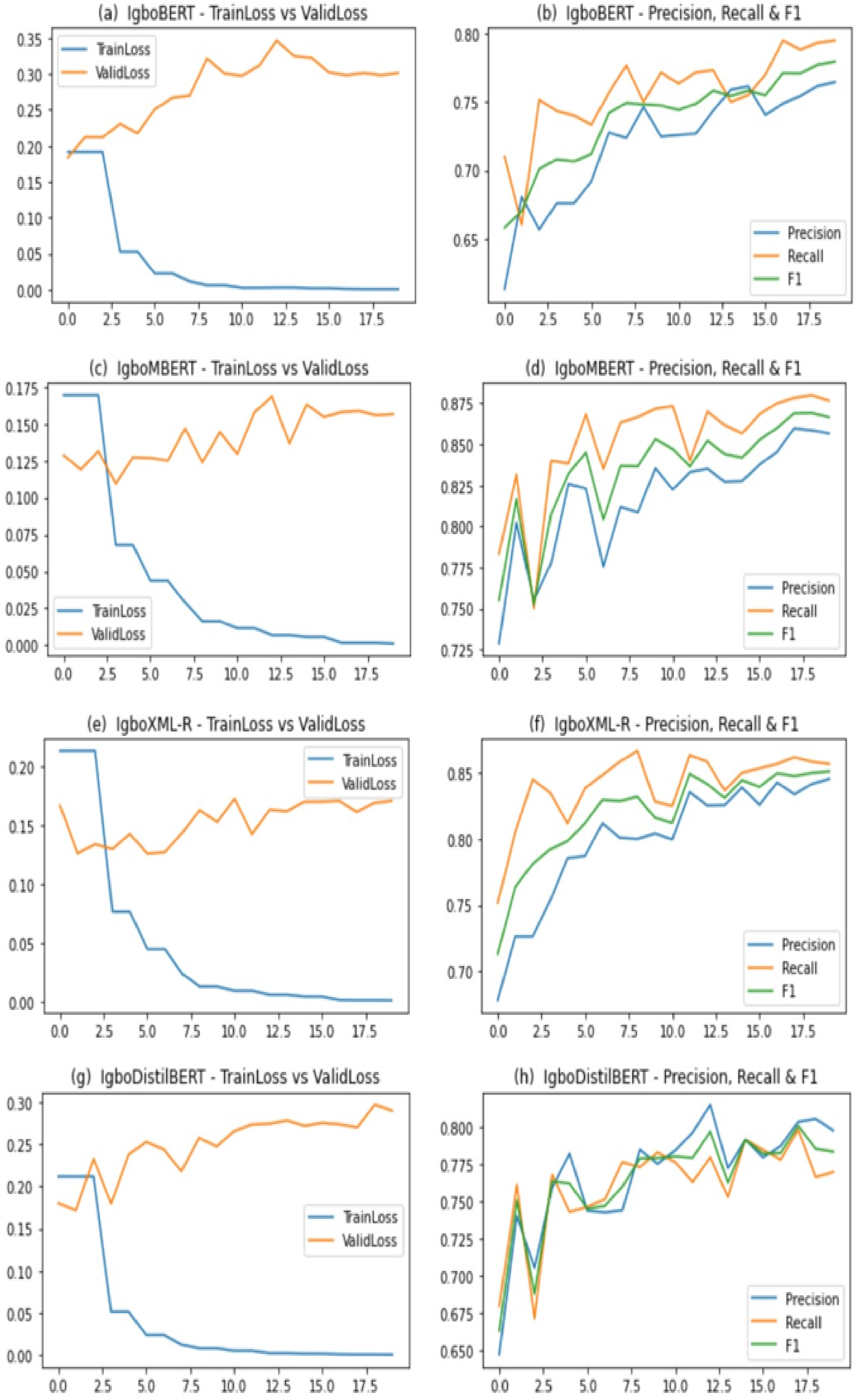

Figure 2: The TrainLoss vs ValidationLoss; Precision, Recall and F1-score of IgboBERT, IgboDistillBERT, IgbomBERT, IgboXML-R at learning rate 1e-4.

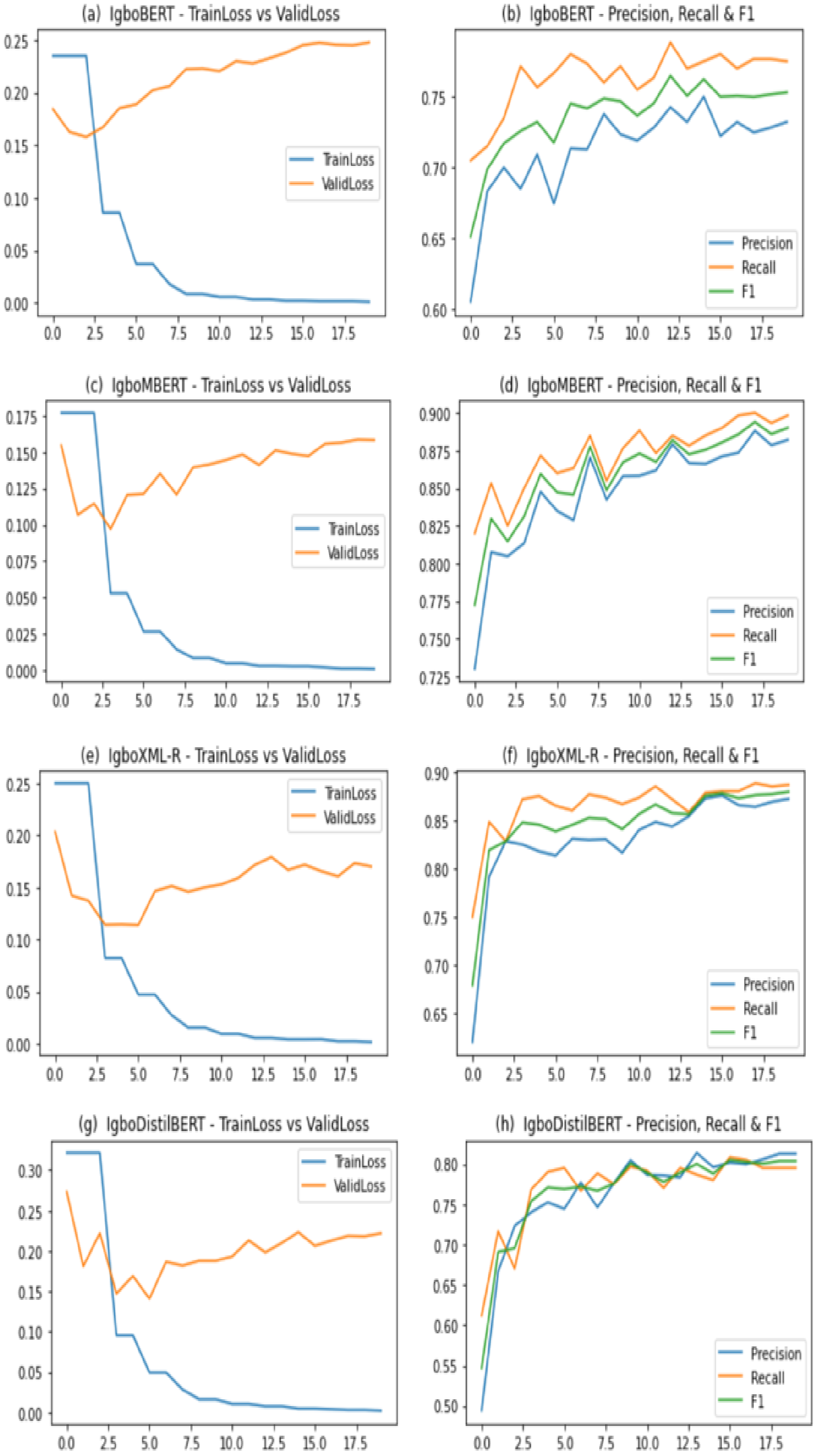

Figure 3: The TrainLoss vs ValidationLoss; Precision, Recall and F1-score of IgboBERT, IgboDistillBERT, IgbomBERT, IgboXML-R at learning rate 2e-5.

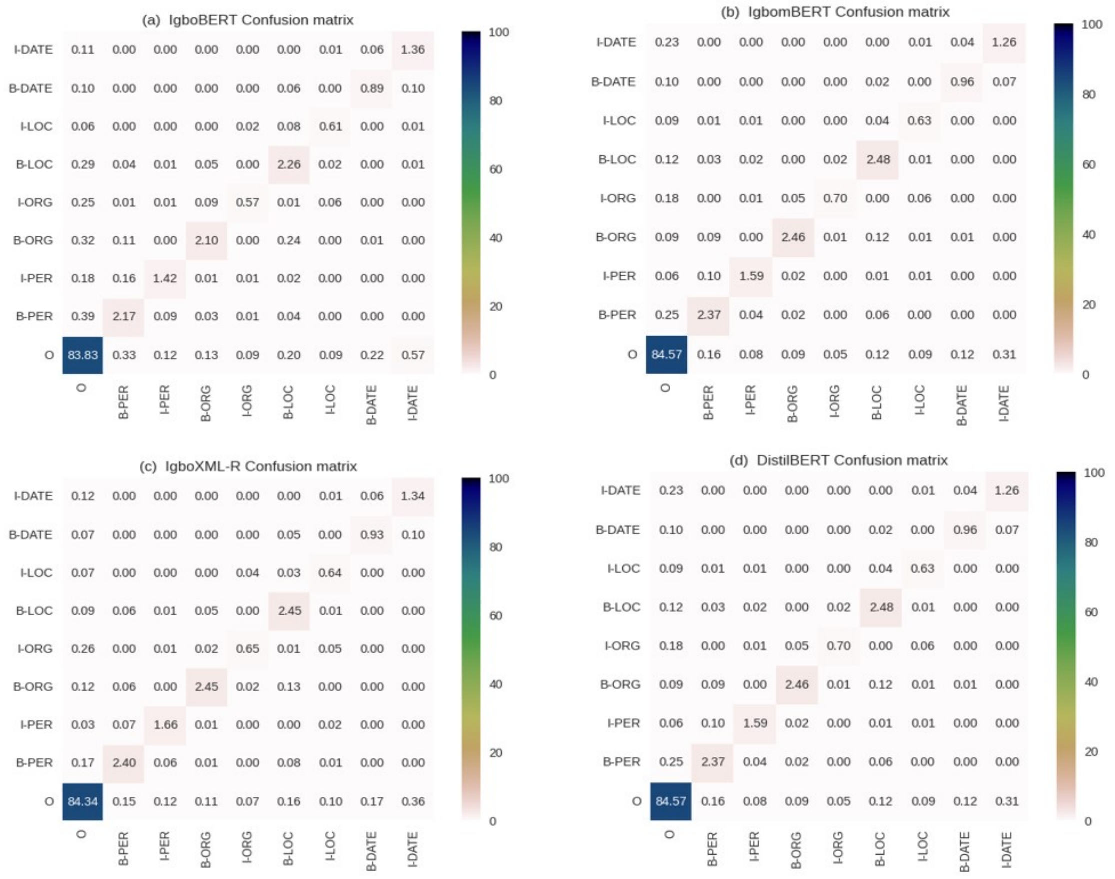

Figure 4: The confusion matrix of IgboBERT, IgbomBERT, IgboXML-R, IgboDistillBERT at learning rate 2e-5.

that it was trained on a huge model of 84M parameters and it was pre-trained on a relatively small raw data when compared to the millions of data used to pre-train mBERT, XML-R and DistilBERT. This resulted in no convergence in the training vs. validation loss (over-fitting). Given that IgboBERT achieves an F1 of 77.94 with such small data, the introduction of more data for fine-tuning may well improve the performance further. We will be handling the issue of over-fitting among other research directions by automatically creating further IgboNER datasets and also use gazetteers to increase our coverage in further studies. The code and model have been released in a GitHub repository[13] to facilitate future research in Igbo NLP and African NLP at large.

8. Acknowledgements

We acknowledge the support of Tertiary Education Trust Fund (TETFund) in Nigeria, and would like to thank the anonymous reviewers for their useful comments and for reading the draft. We also acknowledge JW, Igbo-Radio and Kaoditaa who gave us permission to use data from their websites.

9. Ethical Approval

Approval was obtained from the Faculty of Science and Technology Research Ethics Committee (FSTREC), Lancaster University with Reference: FST21020.

10. Bibliographical References

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., et al. (2021). Masakhaner: Named entity recognition for african languages. *arXiv preprint arXiv:2103.11811*.

Alabi, J., Amponsah-Kaakyire, K., Adelani, D., and España-Bonet, C. (2020). Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France, May. European Language Resources Association.

Babych, B. and Hartley, A. (2003). Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

[13] https://github.com/Chiamakac/IgboNER-Models

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2019). Ethnologue: Languages of the world . sil international.

Eberhard, D. M., Simons, G. F., and (eds.), C. D. F. (2020). Ethnologue: Languages of the world. twenty-third edition.

Ezeani, I., Hepple, M., and Onyenwe, I. (2016). Automatic restoration of diacritics for igbo language. In *International Conference on Text, Speech, and Dialogue*, pages 198–205. Springer.

Ezeani, I., Rayson, P., Onyenwe, I., Uchechukwu, C., and Hepple, M. (2020). Igbo-english machine translation: An evaluation benchmark.

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

Grishman, R. and Sundheim, B. (1995). Message understanding conference 6. In *Proceedings of the 16th conference on Computational linguistics*, volume 1, page 466.

Guo, J., Xu, G., Cheng, X., and Li, H. (2009). Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.

Hedderich, M. A., Adelani, D., Zhu, D., Alabi, J., Markus, U., and Klakow, D. (2020). Transfer learning and distant supervision for multilingual transformer models: A study on african languages. *arXiv preprint arXiv:2010.03179*.

Hedderich, M. A., Lange, L., and Klakow, D. (2021). ANEA: distant supervision for low-resource named entity recognition. *CoRR*, abs/2102.13129.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California, June. Association for Computational Linguistics.

Lin, Y., Costello, C., Zhang, B., Lu, D., Ji, H., Mayfield, J., and McNamee, P. (2018). Platforms for non-speakers annotating names in any language. In *Proceedings of ACL 2018, System Demonstrations*, pages 1–6, Melbourne, Australia, July. Association for Computational Linguistics.

Masakhane, Adelani, D., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., Mayhew, S., Azime, I., Muhammad, S., Emezue, C., Nakatumba-Nabende, J., Ogayo, P., Anuoluwapo, A., Gitau, C., Mbaye, D., Alabi, J., Yimam, S., Gwadabe, T., Ezeani, I., Niyongabo, R., Mukiibi, J., Otiende, V., Orife, I., David, D., Ngom, S., Adewumi, T., Rayson, P., Adeyemi, M., Muriuki, G., Anebi, E., Chukwuneke, C., Odu, N., Wairagala, E., Oyerinde, S., Siro, C., Bateesa, T., Oloyede, T., Wambui, Y., Akinode, V., Nabagereka, D., Katusiime, M., Awokoya, A., MBOUP, M., Gebreyohannes, D., Tilaye, H., and Nwaike, K. (2021). Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, October.

Mollá, D., Van Zaanen, M., Smith, D., et al. (2006). Named entity recognition for question answering.

Nobata, C., Sekine, S., Isahara, H., and Grishman, R. (2002). Summarization system integrated with named entity tagging and ie pattern discovery. In *LREC*.

Onyenwe, I. E. and Hepple, M. (2016). Predicting Morphologically-Complex Unknown Words in Igbo. In *international conference on text, speech, and dialogue*, pages 206–214. Springer.

Onyenwe, I., Uchechukwu, C., and Hepple, M. (2014). Part-of-speech Tagset and Corpus Development for Igbo, an African Language. In *Proceedings of LAW VIII-The 8th Linguistic Annotation Workshop*, pages 93–98.

Onyenwe, I. E., Hepple, M., Chinedu, U., and Ezeani, I. (2018). A Basic Language Resource Kit Implementation for the Igbo NLP Project. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(2):1–23.

Oraka, L. N. (1983). *The foundations of Igbo studies: A short history of the study of Igbo language and culture*. University Publishing Company.

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, Colorado, June. Association for Computational Linguistics.

Ruder, S., Søgaard, A., and Vulić, I. (2019). Unsupervised cross-lingual representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 31–38.

Sanh, V., Debut, L., Chaumond, J., and Wolf,

T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Tsygankova, T., Marini, F., Mayhew, S., and Roth, D. (2021). Building low-resource ner models using non-speaker annotations. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 62–69.

Uchechukwu, C. (2008). African language data processing: The example of the Igbo language. In *10th International pragmatics conference, Data processing in African languages*.