




# Improving Neural Machine Translation for Low Resource Languages Using Mixed Training: The Case of Ethiopian Languages

Atnafu Lambebo Tonja<sup>(✉)</sup>, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),  
Mexico City, Mexico

atnafu.lambebo@wsu.edu.et, kolesolga@gmail.com,  
{mariff2021,gelbukh,sidorov}@cic.ipn.mx

**Abstract.** Neural Machine Translation (NMT) has shown improvement for high-resource languages, but there is still a problem with low-resource languages as NMT performs well on huge parallel data available for high-resource languages. In spite of many proposals to solve the problem of low-resource languages, it continues to be a difficult challenge. The issue becomes even more complicated when few resources cover only one domain. In our attempt to combat this issue, we propose a new approach to improve NMT for low-resource languages. The proposed approach using the transformer model shows 5.3, 5.0, and 3.7 BLEU score improvement for Gamo-English, Gofa-English, and Dawuro-English language pairs, respectively, where Gamo, Gofa, and Dawuro are related low-resource Ethiopian languages. We discuss our contributions and envisage future steps in this challenging research area.

**Keywords:** Machine translation · Low-resource machine translation · Neural machine translation · Ethiopian languages · Mixed training

## 1 Introduction

Natural Language Processing (NLP) is a branch of artificial intelligence which employs computational techniques for the purpose of learning, understanding, and producing human language content [1]. Machine Translation (MT) is one of the widely used NLP applications that carries out the automatic translation from one language to another in order to facilitate communication between people who speak different languages [1]. There are different MT approaches that are being proposed by different researchers and industries to facilitate this task:- from these approaches, Neural Machine Translation (NMT), also known as corpus-based/data-driven machine translation is a current state-of-the-art technique that uses neural networks [2]. NMT is trained on a large corpus of language segments and their respective translations, usually containing hundreds of thousands or even millions of translation units [2].

Researchers have shown that NMT can perform much better than other machine translation models [3]. The quality of NMT as a data-driven approach, massively depends on the quantity, quality, and relevance of the training dataset [2,4]. NMT currently achieved promising results for high-resource languages [4–7], however, it is still inadequate for low-resource conditions [4]. NMT for languages that have low and limited resources is currently one of the research directions in the area of NLP and MT to enable under resource languages to be presented in digital space and to help language speakers to access the current advancement in technologies.

### 1.1 High vs Low Resource Languages

Currently, in the NLP community there is no single way of defining a language as low resource, researchers have proposed various criteria to distinguish high and low resource languages. Joshi et al. [8] created a language resource availability classification based on the amount of existing labeled and unlabeled data. The scale goes from 0 (lowest resources) to 5 (highest resources). One of the details in determining the amount of data is how a data unit is defined. For example, MT data is measured by a number of parallel sentences. The authors also added that high resource languages are characterized by a dominant online presence, it implies massive industrial and government investments in the development of resources and technologies for such languages. On the other hand low resource languages have been suffering from a lack of new language technology designs. When the resources are limited and a little amount of unlabeled data is available, it is very hard to reach a true breakthrough in creating powerful novel methods for language applications.

This paper discusses a new method to improve the performance of NMT for low-resource languages by mixing their data in different scenarios using four languages spoken in Ethiopia as an example. The paper is organized as follows: Sect. 2 describes the previous research related to this study, Sect. 3 gives an overview of the dataset statistics and language description, Sect. 4 explains the methodology adopted in this study, Sect. 5 presents the experimental results and discussion. Finally, Sect. 6 concludes the paper and sheds some light on possible future work.

## 2 Related Work

Mechanization of translation is one of the human beings oldest dream, it became a reality in the 20th century, in the form of computer programs capable of translating a wide variety of texts from one natural language into another [9]. There are different approaches used by researchers for machine translation, some of these are rule-based MT [10,11], statistical MT [12,13], hybrid MT [14] and neural MT [15–17]. Among these approaches neural MT is most efficient current state-of-the-art [2,3] trained on huge datasets containing sentences in a source language and their equivalent target language translations. Basically, NMT takes

advantage of huge translation memories with hundreds of thousands or even millions of translation units [16]. However, NMT for low-resource languages [18] still under-performs due to scarcity of parallel datasets.

A lot of research has been done to solve machine translation problems in low-resource languages, most of them are focused on training a model on high-resource language data and applying transfer learning methodology to low-resource texts. The model trained on a high-resource language pair is called the parent model, then some of the learned parameters are transferred to low-resource pairs (the child model) to initialize and constrain training. During the experiments in [19], the authors used French as the parent source language, and Hausa, Turkish, Uzbek, and Urdu with English as the target language to build the child model. The authors used 300 million English tokens to train French-English parent model and 1.8 million tokens for each of the low-resource languages. As result of transfer learning, they improved the baseline Syntax Based Machine Translation (SBMT) model by an average of 5.6 BLEU on the four low-resource language pairs still leaving room for improvement by selecting parent languages that are more similar to child languages. Concerning the context of our work there are two very important details to note in [19]. First, the researchers used 1.8 million tokens for their selected low-resource languages, however, comparing to our case of Ethiopian languages, such a dataset can be called big. Second, the authors performed experiments in only one direction, i.e., from a low-resource language to English, but not vice versa.

Feng et al. [20] also presented a transfer learning method which improved the BLEU score of the low-resource machine translation. They used an encoder-decoder framework with attention mechanism to train one NMT model for a high resource (French-English) pair, then employed some parameters of the trained model to initialize another NMT model for the Vietnamese-English pair with less parallel corpora compared to the parent French-English model. For the French-English case a 2 million parallel dataset to train NMT as the parent model was used. For the low-resource Vietnamese and Mongolian languages, 133K and 67K sentence pairs were used, respectively. On the Vietnamese-English translation task, their model improved the translation quality by an average of 1.55 BLEU score. Besides, they also got an increase of 0.99 BLEU score translating from Mongolian to Chinese [20].

Slim et al. [21] worked on a transfer learning approach for low-resource NMT applied to the Algerian Arabic dialect. The authors used a fine-tuning transfer learning strategy to transfer knowledge from the parent model(multi-dialects Arabic) to the child model(Algerian dialect). They used a 52K dataset to train the parent model and 12.8K parallel dataset to train the child model using Seq2Seq and attentional-Seq2Seq techniques. The researchers compared the performance of these techniques before and after transfer learning showing that transfer learning improves the BLEU score for the Seq2Seq model from 0.3 to more than 34.56, and for the Attentional-Seq2Seq model from 16.5 to 35.87 for Algerian-English translation.

The above transfer learning approaches showed improvements in low-resource languages but they still use a high-resource language as parent language. In this paper we explore the way of improving NMT performance for low-resource languages without using high-resource language data.

### 3 Dataset

For our experiments we used four related Ethiopian languages spoken in the Southern part of Ethiopia, which are categorized under the same language group and family. Ethiopia is a linguistically and ethnically heterogeneous country with more 80 officially recognized languages [22]. Four languages chosen for our study, namely Wolaita, Gamo, Gofa and Dawuro; are grouped under the Omotic language group which is one of the six language families within the Afro-Asiatic phylum, predominantly spoken in the region between the lakes of the Southern rift valley and the Omo River [23]. The languages in the large Omotic group are classified together as an agentic unit because their phonology, grammar and lexicon are quite close.

Parallel datasets used in this study are borrowed from the NMT research conducted by [24], they belong to the religious domain available in digital form. Table 1 shows parallel dataset distribution of the four languages in this study. It can be observed in Table 1, that English has less type count than Ethiopian languages. This demonstrates the morphological complexity of Ethiopian languages. The number of sentences in the Wolaita-English pair is bigger than in the others, due to the fact that, the Wolaita-English parallel dataset contains both the Old and New-Testaments of the Bible, whereas the rest contain only the New Testament of the Bible.

**Table 1.** Parallel dataset distribution

Languages	Sentences	Tokens	Types	Average words in a sentence
English	26,943	703,122	12,131	26
Wolaita		469,851	42,049	17
English	7,866	177,410	11,078	23
Gamo		125,509	23,589	16
English	7,928	175,727	8,769	22
Gofa		119,289	25,301	15
English	7,804	207,954	4,368	27
Dawuro		126,734	17,392	16

### 4 Methodology

This section describes the proposed model and data processing used for our proposed approach. For our NMT approach we employed the transformer model,

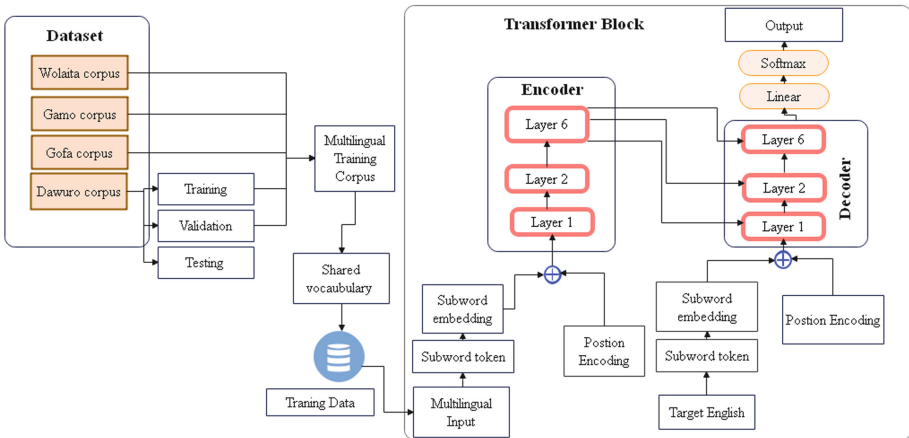
the current state-of-the-art deep learning model occupied primarily in the fields of natural language processing after its first introduction in [25]. Like Recurrent Neural Networks (RNNs), transformers are designed to handle sequential input data, e.g., natural language texts, for tasks such as translation and text summarization. However, unlike RNNs, transformers do not necessarily process the data in order. Rather, the attention mechanism provides context for any token position in the input sequence [25].

#### 4.1 Data Pre-processing

Before training the NMT model with the datasets described previously, we pre-processed them to facilitate training by converting text data into a format suitable for our model. This phase includes removing duplicate entries, characters that are not in Latin scripts, removing digits, and converting each sentence into lower case.

#### 4.2 Model

Our NMT model based on the transformer architecture contains both Encoder and Decoder blocks with six layers, input embedding(subword embedding), output embedding (subword embedding), positional encoding, linear classifier, and the final Softmax layer for output probabilities. Figure 1 shows the proposed transformer-based model architecture for low-resource NMT. As shown in Fig. 1, our proposed model has a dataset block and a transformer block. The dataset block contains datasets of four languages. To mix the data, we developed the procedure which we will explain here for the Dawuro language as an example. We take the data of the three datasets except Dawuro. Then we split the Dawuro-English dataset into training, validation, and test subsets, combine the training



**Fig. 1.** Proposed transformer-based model for low-resource NMT

and validation subsets with the data of the three(Gamo, Gofa and Wolaita) languages for training and further on apply the Dawuro-English test subset to evaluate the model performance. Such procedure is used to train NMT for the other three languages. The source and target parallel sentences are converted into subword tokens using Byte Pair Encoding (BPE) representation [26] and further to subword embedding as input to the positional encoder block.

## 5 Experiments and Results

This section describes the experimental settings, dataset split strategies and performance of the models and comparison with the previous studies.

### 5.1 Experiments

We trained our translation models using an open source ecosystem for neural machine translation and neural sequence learning toolkit called OpenNMT [27] with tensorflow version and transformer [25]. To conduct the study we used OpenNMT-tf [27] in Google colab pro + [28] with Graphical processing Unit(GPU). We used transformer [25] and BPE [26] subword tokenization which is a simple form of data compression algorithm in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur in that data. The BPE representation was chosen in order to remove vocabulary overlap during dataset combinations. Table 2 shows the parameters used to train our proposed model.

**Table 2.** Parameters used for training and evaluation of the proposed NMT model

Parameters	Values
Encoder layers	6
Decoder layers	6
Model-dim	512
Learning rate	0.0001
Drop out	0.3
Optimizer	Adam
Batch size	3072
Batch type	Tokens

### 5.2 Dataset Split

In order to carry out our experiments we divided the datasets into training, validation and test set. As shown in the Table 3 except for Experiment(1) we used Wolaita-English dataset without splitting in other experiments. For Experiment

(1) (2) and (4) we used the same splitting method for three (Gamo, Gofa, and Dawuro) language pairs. For Experiment (3) we used Wolaita- English dataset for training and three language pairs for validation and testing.

**Table 3.** Training, validation and test set split for the experiments. The numbers indicates the amount of parallel sentences used in each split

Experiments	Language Pair	Training set	Validation set	Test set
Exp-1	Wolaita-English	21,555	5,388	-
	Dawuro-English	4,996	1,248	1,560
	Gamo-English	5,035	1,258	1,573
Transfer Learning	Gofa-English	5,077	1,267	1,584
Exp-2	Wolaita-English	26,943	-	-
	Dawuro-English	4,996	1,248	1,560
	Gamo-English	5,035	1,258	1,573
	Gofa-English	5,077	1,267	1,584
Exp-3	Wolaita-English	26,943	-	-
	Dawuro-English	-	1,561	6,243
	Gamo-English	-	1,574	6,293
	Gofa-English	-	1,584	6,335
Exp-4	Wolaita-English	26,943	-	-
	Dawuro-English	4,996	1,248	1,560
	Gamo-English	5,035	1,258	1,573
	Gofa-English	5,077	1,584	1,267

### 5.3 Results

We conducted four experiments for low-resource NMT. In **Experiment (1)** we trained the model using a transfer learning approach, for this experiment we first trained the parent (Wolaita - English) model on the Wolaita-English parallel dataset and fine-tuned it on the resting three language pairs. In **Experiment (2)** we trained the NMT model by combining the Wolaita - English dataset with one of the resting language pairs. In **Experiment (3)** we trained the model by combining the Wolaita -English dataset with two language pairs, validated and tested the performance of the model with the unused language pair. In **Experiment (4)** we trained the model by combining the Wolaita - English dataset with three language pairs and tested the performance on each of the language pairs.

As shown in Table 4, using English as a target language, combining two or more low-resource languages(**Experiment 2 & 4**) gives better results than using low-resource languages as parent languages for the transfer learning approach (**Experiment 1**) and testing on unused language pairs (**Experiment 3**). This shows the possibility of improving NMT performance for low-resource languages by:-

**Table 4.** Low-resource NMT experimental results using English as the target language

Language	Exp1			Exp2			Exp3			Exp4		
	BLEU	TER	chrF2++	BLEU	TER	chrF2++	BLEU	TER	chrF2++	BLEU	TER	chrF2++
Gamo-English	1.4	106.2	20.4	6.5	88	22.8	4.8	94.4	22.2	9.4	83.6	26.9
Dawuro-English	3.9	92.5	21.5	6.2	87.2	23.1	2.2	101.7	20.9	7.3	83.5	26.5
Gofa-English	5.6	99.3	22.4	7.2	86.8	25.1	5.4	93.1	25.6	9.5	83.3	27.1

**Table 5.** Low-resource NMT experimental results using English as source languages

Language	Exp1			Exp2			Exp3			Exp4		
	BLEU	TER	chrF2++	BLEU	TER	chrF2++	BLEU	TER	chrF2++	BLEU	TER	chrF2++
English - Gamo	-	-	-	2.0	100.7	21	-	-	-	1.5	103.2	20.7
English - Dawuro	-	-	-	2.6	97.7	21.4	-	-	-	1.8	101.4	21.2
English - Gofa	-	-	-	2.9	97.5	22.3	-	-	-	2.1	100.5	21.2

- Combining two related languages, one with more resources and the other with fewer resources for training and using less resource language for validation and testing.
- Combining more than two related languages one with more resources and the others with fewer resources for training and using less-resource language for testing and validation.

As shown in Table 5 using English as a source language gives poor results compared to using English as the target language in Table 4, because the model favors the English data over the Ethiopian data due to the morphological richness and complexity of the Ethiopian languages. In addition to this, when English is used as the source language, the translation is challenged by many-to-one alignment.

**Table 6.** Comparison of the proposed approach with previous studies

Methods	Languages	BLEU score	
		English-*	*-English
Tonja et al. [24]	Gamo*	2.2	4.1
	Gofa*	2.4	4.5
	Dawuro*	2.1	3.6
Yigezu et al. [29]	Gamo*	-	3.4
	Gofa*	-	4.5
	Dawuro*	-	2.5
Our proposed approach(Exp2)	Gamo*	2.0	<b>6.5</b>
	Gofa*	<b>2.9</b>	<b>7.2</b>
	Dawuro*	<b>2.6</b>	<b>6.2</b>
Our proposed approach(Exp4)	Gamo*	1.5	<b>9.4</b>
	Gofa*	2.1	<b>9.5</b>
	Dawuro*	1.8	<b>7.3</b>



In Table 6 and Fig. 2, we compared our proposed approaches with previous studies [24, 29] that used the same datasets and languages. It can be seen that the results of our **Experiments 2** and **4** show improvement over previous works on the same language pairs with English as a target language. This evidences that using a combination of more than two related low-resource languages for training improves the performance of NMT for low-resource languages without using high-resource languages in one direction. Besides, a combination of two related low-resource languages improves translation in both directions.

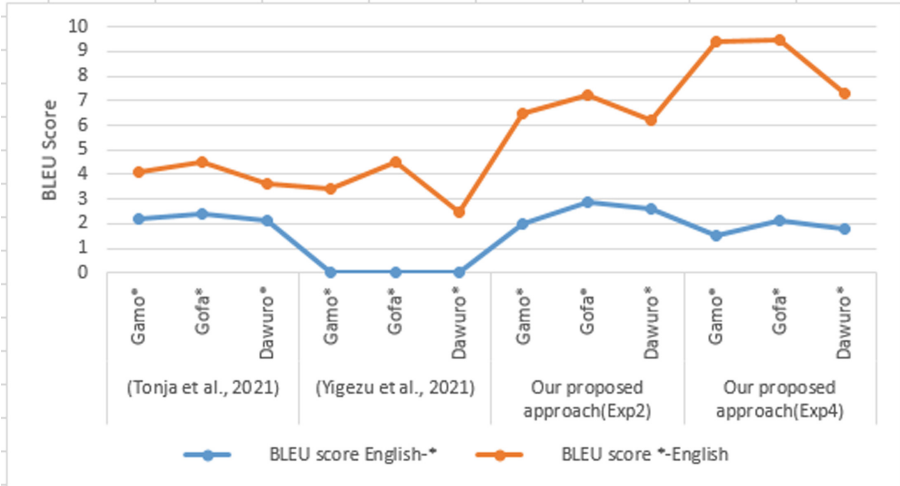


Fig. 2. Comparison of the proposed approach with previous research works

## 6 Conclusions and Future Work

In this paper, we proposed and discussed a new approach to improve neural machine translation for low-resource languages using a transformer-based model. Combining two or more low-resource languages for training and validating, then testing the performance on another language shows result improvements when English is used as the target language. Our proposed model showed better results compared to previous experiments in the same languages and datasets without using high-resource languages.

In future work, we will apply the proposed approach for other related low-resource languages and compare the performance of the proposed approach with other suggested NMT approaches for low-resource languages. Also, we plan to add more domains and investigate the effect of domain for low-resource languages in machine translation.

**Acknowledgments.** The work was done with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico, grants 20220852,

20220859, and 20221627 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

## References

1. Julia, H., Manning, C.D.: Advances in natural language processing. *Science* **349**(6245), 261–266 (2015)
2. Forcada, M.L.: Making sense of neural machine translation. *Transl. Spaces* **6**(2), 291–309 (2017)
3. Mohamed, S.A., Elsayed, A.A., Hassan, Y.F., Abdou, M.A.: Neural machine translation: past, present, and future. *Neural Comput. Appl.* **33**(23), 15919–15931 (2021). <https://doi.org/10.1007/s00521-021-06268-0>
4. Benyamin, A., Dorr, B.J.: Augmenting neural machine translation through round-trip training approach. *Open Comput. Sci.* **9**(1), 268–278 (2019)
5. Alexandre, B., Kim, Z.M., Nikoulina, V., Park, E.L., Gallé, M.: A multilingual neural machine translation model for biomedical data. arXiv preprint [arXiv:2008.02878](https://arxiv.org/abs/2008.02878) (2020)
6. Markus, F., Firat, O.: Complete multilingual neural machine translation. arXiv preprint [arXiv:2010.10239](https://arxiv.org/abs/2010.10239) (2020)
7. Khaled, S., Rafea, A., Moneim, A.A., Baraka, H.: Machine translation of English noun phrases into Arabic. *Int. J. Comput. Process. Oriental Lang.* **17**(02), 121–134 (2004)
8. Pratik, J., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.: The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint [arXiv:2004.09095](https://arxiv.org/abs/2004.09095) (2020)
9. Hasibuan, Z.: A comparative study between human translation and machine translation as an interdisciplinary research. *J. Eng. Teach. Learn. Issue.* **3**(2), 115–130 (2020)
10. Dubey, P.: Study and development of machine translation system from Hindi language to Dogri language an important tool to bridge the digital divide (2008)
11. Okpor, M.D.: Machine translation approaches: issues and challenges. *Int. J. Comput. Sci. Issue. (IJCSI)* **11**(5), 159 (2014)
12. Lopez, A.: Statistical machine translation. *ACM Comput. Surv. (CSUR)* **40**(3), 1–49 (2008)
13. Philipp, K.: *Statistical Machine Translation*. Cambridge University Press (2009)
14. Sergei, N., Somers, H.L., Wilks, Y.A.: A Framework of a Mechanical Translation between Japanese and English by Analogy Principle, pp. 351–354 (2003)
15. Philipp, K.: Neural machine translation. arXiv preprint [arXiv:1709.07809](https://arxiv.org/abs/1709.07809) (2017)
16. Stahlberg, F.: Neural machine translation: a review. *J. Artif. Intell. Res.* **69**, 343–418 (2020)
17. Wu, Y., et al.: Google’s neural machine translation system: bridging the gap between human and machine translation. arXiv preprint [arXiv:1609.08144](https://arxiv.org/abs/1609.08144) (2016)
18. Robert, Ö., Tiedemann, J.: Neural machine translation for low-resource languages. arXiv preprint [arXiv:1708.05729](https://arxiv.org/abs/1708.05729) (2017)
19. Barret, Z., Yuret, D., May, J., Knight, K.: Transfer learning for low-resource neural machine translation. arXiv preprint [arXiv:1604.02201](https://arxiv.org/abs/1604.02201) (2016)

20. Tao, F., Li, M., Chen, L.: Low-resource neural machine translation with transfer learning. In: LREC 2018 Workshop, p. 30 (2018)
21. Amel, S., Melouah, a., Faghihi, u., Sahib, k.: Improving neural machine translation for low resource Algerian dialect by transductive transfer learning strategy. Arab. J. Sci. Eng. **47**, 10411–10418 (2022)
22. Hirut, W.: Language planning challenged by identity contestation in a multilingual setting: the case of gamo. Oslo Stud. Lang. **8**(1) (2016)
23. Azeb, A.: The Omotic Language Family. Cambridge University Press (2017)
24. Atnafu Lambebo, T., Woldeyohannis, M.M., Yigezu, M.G.: A parallel corpora for bi-directional neural machine translation for low resourced Ethiopian languages. In: 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pp. 71–76. IEEE (2021)
25. Ashish, V., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
26. Gage, P.: A new algorithm for data compression. C Users J. **12**(2), 23–38 (1994)
27. Guillaume, K., Kim, Y., Deng, Y., Senellart, J., Rush, A.M.: OpenNMT: open-source toolkit for neural machine translation. arXiv preprint [arXiv:1701.02810](https://arxiv.org/abs/1701.02810) (2017)
28. Michael, C., Bragança, L., Paranaiba Vilela Neto, O., Nacif, J.A., Ferreira R.: Google colab cad4u: hands-on cloud laboratories for digital design. In: 2021 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5. IEEE (2021)
29. Yigezu, M.G., Woldeyohannis M.M., Tonja, A.L.: Multilingual neural machine translation for low resourced languages: Ometo-English. In: 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), pp. 89–94 (2021). <https://doi.org/10.1109/ICT4DA53266.2021.9671270>