applied sciences

MDPI

Article

Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data

Atnafu Lambebo Tonja, Olga Kolesnikova *, Alexander Gelbukh and Grigori Sidorov

Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City 07738, Mexico
* Correspondence: kolesnikova@cic.ipn.mx

Abstract: Despite the many proposals to solve the neural machine translation (NMT) problem of low-resource languages, it continues to be difficult. The issue becomes even more complicated when few resources cover only a single domain. In this paper, we discuss the applicability of a source-side monolingual dataset of low-resource languages to improve the NMT system for such languages. In our experiments, we used Wolaytta–English translation as a low-resource language. We discuss the use of self-learning and fine-tuning approaches to improve the NMT system for Wolaytta–English translation using both authentic and synthetic datasets. The self-learning approach showed +2.7 and +2.4 BLEU score improvements for Wolaytta–English and English–Wolaytta translations, respectively, over the best-performing baseline model. Further fine-tuning the best-performing self-learning model showed +1.2 and +0.6 BLEU score improvements for Wolaytta–English and English–Wolaytta translations, respectively. We reflect on our contributions and plan for the future of this difficult field of study.

Keywords: Wolaytta–English NMT; English–Wolaytta NMT; low-resource NMT; self-learning; neural machine translation; monolingual data for low-resource languages; low-resource NMT

Citation: Tonja, A.L.; Kolesnikova, O.; Gelbukh, A.; Sidorov, G. Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data. *Appl. Sci.* **2023**, *13*, 1201. https://doi.org/10.3390/app13021201

Academic Editor: Douglas O'Shaughnessy

Received: 9 November 2022
Revised: 7 January 2023
Accepted: 12 January 2023
Published: 16 January 2023

1. Introduction

Machine Translation (MT) is a conventional name for computerized systems that produce translation from one natural language to another, with or without human intervention [1]. As computational activities become more mainstream and the internet opens to a wider multilingual and global community, research and development in MT continue to grow at a rapid rate [2]. Linguists and computer scientists have been working together on MT for the past few years and have made a lot of progress on this hard task. There have been numerous MT approaches developed, including rule-based and example-based and statistical machine translation (SMT), and neural machine translation (NMT) approaches [1]. Recent advances in deep learning have led to the dominance of neural network-based methods in various subfields of artificial intelligence (AI), including NMT [3]. NMT is the current state-of-the-art approach in MT, and it learns from a huge dataset of source language sentences and their translations in the target language [4]. Deep neural methods, including NMT, need a lot of data and cannot be trained well in environments with few resources. Even though MT has come a long way in the last ten years, it has mostly been used for high-resource languages because neural networks need large corpora to train well.

Therefore, the success of NMT models is heavily dependent on the availability of extended parallel data, which can only be collected for a limited number of language pairs [5]. High translation quality has been achieved for high-resource language pairings, such as English–Spanish [6–8], English–French [6,8,9], English–Russian [6,8,9], and English–Portuguese [8]. NMT systems perform poorly in low-resource environments due to the lack of enough training data for those languages [10]. Low-resource languages are not as well represented in digital spaces as high-resource languages, which makes it hard for people who speak these languages to use the latest technologies in their daily lives, including

effective NMT systems. In order to address this problem, researchers have proposed several solutions, including multilingual NMT [11], transfer learning [12], exploiting related languages [13], multimodal NMT [14], data augmentation [15], filtered pseudo-parallel datasets [16], and meta-learning [17].

However, most of the suggested approaches used high-resource languages for their experiments, and to the best of our knowledge, no research has been performed using source-side monolingual data of low-resource languages to improve NMT. Using Wolaytta–English translation as a case study, we propose an efficient method for using such data to improve the NMT performance in low-resource languages. We collected 40 k Wolaytta monolingual data from several sources, such as Wolaytta Fana Radio, Wolaytta Wogeta Radio, and Wolaytta textbooks and used them along with a Wolaytta–English parallel dataset in two experiments: (i) training a model on the available Wolaytta–English parallel dataset (our baseline model) and (ii) training the selected model from the baseline on the combination of the authentic dataset and synthetic dataset with a self-learning and fine-tuning approach following the procedure in [18].

Our objective is to answer the following research questions:

- Will relying solely on source-side monolingual data improve the NMT performance for a low-resource language?
- Will monolingual data from a low-resource language improve the performance of NMT when English is used as the source language?

To respond to the above questions and to explore the NMT performance in a low-resource language, we focus on NMT for the Wolaytta–English language pair because this language pair has parallel data only from the religion domain, and this fact also serves as an illustration of how difficult it is to compile a parallel dataset for a low-resource language. Wolaytta is a language spoken in the southern part of Ethiopia, with few or no digital resources [19]; a more detailed description can be found in Section 3.

In this paper, we investigate NMT training from a dataset perspective for low-resourced languages. We specifically examine the effects of using low-resource language source-side monolingual data on improving the NMT and the most effective ways to use such data. As a result, the contributions of our work are the following:

- We thoroughly investigated the usability of a source-side monolingual dataset for low-resource languages to improve NMT.
- We found a good way to combine a synthetic parallel dataset with an authentic parallel dataset when training an NMT model for low-resource languages.
- We developed a good training method for fine-tuning NMT models that were trained on a mixed set of synthetic and authentic datasets.
- We made our training scripts and the monolingual and parallel datasets available for public use. Specifically, the datasets can be used in the future as a benchmark for Wolaytta–English machine translation tasks.

The rest of the article is organized as follows: Section 2 describes previous research related to this study; Section 3 describes the Wolaytta language; Section 4 gives some statistics of our dataset; Section 5 explains our proposed methodology; Section 6 presents the experimental results, and Section 7 includes a detailed analysis of our results. Finally, Section 8 concludes the paper and sheds some light on possible future work.

2. Related Work

Researchers have recently come up with a lot of ideas for how to improve NMT for languages with few resources. One way to improve the NMT system, especially for languages with few resources, is to use monolingual data and parallel datasets. In this section, we explore related studies performed for different languages using monolingual datasets as an additional source to improve NMT systems for low-resource languages. Laskar et al. [20] proposed utilizing monolingual data via pretrained word embeddings in transformer model-based neural machine translation to tackle the parallel dataset prob-

lem in bidirectional Tamil–Telugu NMT. The authors used GloVe [21] to pre-train word embedding on the monolingual corpora and used them in the transformer model during the training process. Their model achieved a BLEU score of 4.05 for both Tamil–Telugu and Telugu–Tamil translations, respectively. Marie et al. [22] proposed a new method to generate large synthetic parallel data by leveraging very small monolingual data in a specific domain. They fine-tuned a pre-trained GPT-2 model on small in-domain monolingual data and used the resulting model to generate a large amount of synthetic in-domain monolingual data. Then, they performed a back translation to generate synthetic in-domain parallel data. Their results on the English–German, English–French, and English–Japan pairs and five domains showed improvements in BLEU for all configurations when using synthetic data to train NMT. Tars et al. [23] proposed a multilingual training approach, which can be improved by leveraging monolingual data to create synthetic bilingual corpora using the back translation method to improve low-resource machine translation. Their multilingual learning strategy and synthetic corpora increased the translation quality for language pairs from the Estonian and Finnish geographical regions.

Sennrich et al. [24] utilized monolingual data to improve the performance of the NMT model for the English–German and Turkish–English language pairs. They created synthetic parallel data by translating target language monolingual data into the source language. Then, their initial target–source MT system was trained on the available parallel data. After training, the system translated the monolingual dataset from the source language into the target language to obtain the back-translated data. Next, Sennrich et al. [24] trained the final source–target NMT system by mixing the back-translated data with the original parallel data. Finally, the researchers achieved significant gains on the WMT-15 English–German task with a +2.8–3.7 BLEU improvement and on the low-resource IWSLT-14 Turkish–English task with a +2.1–3.4 BLEU improvement. Jiao et al. [25] proposed self-training sampling with monolingual data uncertainty for NMT to improve the sampling procedure by selecting the most informative monolingual sentences to complement the parallel data. Using the bilingual lexicon generated from the parallel data, they calculated the uncertainty of monolingual sentences. They suggested that emphasizing the learning of uncertain monolingual sentences improves the translation quality of high-uncertainty sentences and benefits the prediction of low-frequency words on the target side. The results of their experiments on large WMT English–German and English–Chinese datasets show that the proposed method improved the performance of NMT.

Dione et al. [26] utilized subword segmentation, back translation, and the copied corpus methods to improve the NMT performance of bidirectional Wolof–French translation. When back translation and the copied corpus were used together, the quality of the translation from Wolof to French showed improvement in both directions. Pham [27] proposed the use of the Google Translate application as a reverse MT method to improve the quality of monolingual texts for back translation in the English–Vietnamese language pair. Compared to the baseline approach for English–Vietnamese NMT, their proposed approach raised the BLEU score by 16.37 points. Ngo et al. [28] proposed extracting vocabulary from the original bilingual corpus's target text and generating artificial translation units by tagging each standard translation with a label. Then, they concatenated the synthetic corpus with the native corpus for training NMT systems. Their proposed approach demonstrated improvements of +1.8 and +1.9 in the BLEU scores for Chinese–Vietnamese and Japanese–Vietnamese translation tasks, respectively.

Table 1 gives a summary of the related works that used monolingual datasets to improve the NMT system for the different language pairs discussed above. The language pairs in the papers have more resources than the language we chose for this study because they are more common in the digital space and have larger parallel corpora. The monolingual datasets used in the reviewed papers were easy to collect, but this is not the case for languages such as Wolaytta.

Table 1. Review of different machine translation systems that used monolingual datasets.

Publication Year	Reference	Language Pairs Used	Parallel Dataset Size	Monolingual Dataset Size	Monolingual Data Used		Approach
					Source-Side	Target-Side	
2021	Laskar et al. [20]	Tamil(ta)–Telugu(te)	43K	Ta (315 M), Te (478 M)	✔	✔	Transformer
2021	Marie et al. [22]	English(en)-German(de), English(en)-French(fr), English(en)-Japan(ja)	En-De (5.1 M), En-Fr (32.7 M), En-Ja (3.9 M)	En (2.98 M)	✖	✔	Transformer
2021	Tars et al. [23]	English(en)-German(de), English(en)-French(fr), English(en)-Japan(ja)	Et-Fi (2.6 M), Fi-Sma (3 k), Et-Vro (30 k), Fi-Sme (109 K), Sme-Sma (3.7 M)	Et (125 K), Fi (125 k), Vro (168 k), Sme (40 K), Sma (60 K)	✔	✔	Transformer
2021	Sennrich et al. [24]	English(en)-German(de), English(en)-Turkish(tr)	En-De (4 M), En-Tr (320 K)	En (118 M)-De (160 M), En (3 M)-Tr	✔	✔	Transformer
2021	Jiao et al. [25]	English(en)-German(de), English(en)-Chinese(zh)	En-De (36.8 M), En-Zh (22.1 M)	En-De (40 M), En-Zh (20 M)	✔	✔	Transformer
2022	Dione et al. [26]	French(fr)-Wolof(wo)	78 k	Wo (35 K), Fr (39 K)	✔	✔	Transformer
2022	Pham [27]	English(en)-Vietnamese(vi)	133 k	300 K Vi	✖	✔	Transformer
2022	Ngo et al. [28]	Chinese(zh)-Vietnamese(vi), Japanese(ja)-Vietnamese(vi)	Zh-Vi (24.4 M), Ja-Vi (24.5 M)	Za (27 M), Ja (27 M)	✔	✖	Transformer

In this paper, we test whether a monolingual dataset of a low-resource language can be used to improve NMT by running experiments on language pairs with small corpora. We explore ways to generate target-side synthertic data for NMT using a small amount of the source-side monolingual dataset and parallel dataset. In our experiments, we followed the training approach of Sennrich et al. [24] to generate synthetic data and train the model on mixed datasets in a self-learning fashion.

3. Overview of the Wolaytta Language

We added this section to show the difference between the Wolaytta language and English in terms of morphology, phonology, writing system, and word order. We hope this helps the reader to understand the languages we are using and how difficult it is to work on low-resource and morphologically complex languages in the area of NMT.

Wolaytta refers to the people, language, and area located in the Wolaytta Zone of southern Ethiopia [29]. The Wolaytta language is one of the languages in the Ometo group, which belongs to the Omotic branch of the Afroasiatic family (or phylum). In addition to the Wolaytta Zone, it is spoken in Addis Ababa city and in other border areas, such as Gamo, Gofa, and Dawuro. The natives refer to their language as *Wolayttattuwa*, although it is also referred to as *Wolaytta doonna* or *Wolaytta Kaalaa*.

It is used as a primary school medium of instruction and as a secondary and high school subject. Currently, the Wolaytta language is offered as a subject in the Bachelor's program at Wolaytta Sodo University, Sodo, Ethiopia (http://www.wsu.edu.et/). In the Wolaytta Zone, this language is used in government offices as the language of work and communication.

3.1. Phonolog

3.1.1. Vowels

Wolaytta has five vowel phonemes, each of them in both long and short variants [30].

3.1.2. Consonants

Wolaytta has the following consonant phonemes: voiceless (p, t, k, P, s, sh, h, nh, and c), voiced (b, d, g, z, zh, j, m, n, r, l, w, and y), and glottalized (P, T, K, C, D, L, M, and N) [29,30].

3.1.3. Writing System

The Wolaytta language employs a Latin-based alphabet with twenty-nine fundamental letters, of which five ('i', 'e', 'a', 'o', and 'u') are vowels, twenty-four ('b', 'c', 'd', 'f', 'g', 'h', 'j', 'k', 'l', 'm', 'n', 'p', 'q', 'r', 's', 't', 'v', 'w', 'x', 'y', 'z', and 'ʔ') are consonants, and seven pair letters fall together (a combination of two consonant characters 'ch', 'dh', 'ny', 'ph', 'sh', 'ts', and 'zh') [27–29]. Numerous textbooks are being produced in the Latin alphabet [31], reflecting the alphabet's widespread use in the mother-tongue of the school system.

3.2. Morphology

Wolaytta is a language with one of the most complex morphological systems and is categorized as an agglutinative type, similarly employing both morphological inflection and derivation like other languages of this kind, leading to a huge number of variations for a single word [30,31]. Wolaytta only depends on suffixes to form different forms of a given word. siiqa ('love'), including siiqa ('love it'), siiqa-asa ('you love'), siiqa-asu ('she loves'), siiqa-da ('fall in love'), siiqa-dasa ('you loved'), siiqa-dii ('do you love'), siiq-idda ('while loving'), siiqa-is ('i love'), siiq-oosona ('they love'), and siiq-ida ('by loving'), are examples of morphological richness. These forms are derived from the root siiqa 'love' by adding the suffixes -asa, -asu, -idda, -is, and -ida. Among the word formation patterns in Wolaytta, the common one is suffixation, in which words are frequently created by combining two or more suffixes [29]. By adding one suffix on top of another, we make a long word that often has as much meaning and grammar as a whole English phrase, clause, or even sentence. For example, (i) he-g-aa-dan ('like') formed by adding three (-g, -aa, and -dan) suffixes to the root word he ('that'), (ii) bochch-enn-aa-dan ('don't touch') formed by adding three (-enn, -aa, and -dan) suffixes to the root word bochcha ('touch'). Because of the intricate nature of its grammatical structure, a single Wolaytta word may have a very wide variety of different meanings when translated into other languages.

3.2.1. Nouns

Wolaytta nouns are divided into four groups based on the endings they take in inflection [29–31]. The nouns ending in '-a' in the absolute case and with the action on their last syllable belong to the first class, for example, aawwa ('sun') and tuma ('truth'). Nouns with their absolute case ending '-iya' and the accent on their penultimate syllable make up the second class, for example, xalliya ('medicine'), ogiya ('road'), and siniya ('cup'). Nouns ending in '-uwa' in their absolute case constitute the third class, for example, ossuwa ('work') and giduwa ('center'). The forth class consists of nouns ending in '-(i)yo' in their absolute case. The letters mainly include terms referring to female living beings, for example, naPP-yo or na''-yo ('the girl') and machas-iyo ('the wife').

3.2.2. Gender

Like many other languages, Wolaytta nouns exhibits two genders: masculine and feminine [29–31]. Nouns belonging to the fourth class are feminine, while the nouns belonging to the other three classes are masculine. The feminine ones differ from the masculine ones by their endings; the former are characterized by the ending '-o' in the absolute case and the latter by the ending '-a' in the absolute case. Additionally, if nouns are used as a subject, they are marked with '-iy' for the feminine and '-ay' for the masculine.

There are a few exceptions, such as possessive pronouns and some demonstratives, which may only have forms for one gender or the other.

3.2.3. Number

According to Wakasa [29] and Hirut [30], the Wolaytta noun system has two numbers: singular and plural. Most nouns have a singular form, and the plural form is formed by adding a suffix to the singular form. Wolaytta forms the plural of nouns by means of a morpheme '-tv', where '-v' represents a final vowel, which changes according to the case inflection of the plural itself. In the absolute case, '-v' corresponds to '-a': *dorssa* ('sheep') and *desha* ('goat'); thus, the plural marker is '-ta': *dorssa-ta* ('sheep') and *desha-ta* ('goats') in that case.

3.3. Articles

Wolaytta does not require articles in front of nouns. The meaning usually represented by the definite article in other languages is rendered by dropping the last vowel of the noun and adding the suffix -ti [32]. For example, 'sheep' is *dorssa* and ('the sheep') is *dorssa –ti*, 'dog' is *kana*, and ('the dogs') is *kana-ti*.

3.4. Syntax

The Wolaytta language employs Subject–Object–Verb (SOV) [31] word order. For example, in the sentence *Bakali ketta kexiis*, *Bakali* ('Bekele') is the subject; *ketta* ('house') is the object, and *kexiis* ('built') is the verb. In English, the typical sentence structure is Subject–Verb–Object (SVO); therefore, the above Wolaytta sentence is translated as "Bekele built a house".

3.5. Punctuation

Except for apostrophes, Wolaytta and English use the same punctuation marks with the same function except for the apostrophe. In Wolaytta, the apostrophe is used to indicate the glottal stop symbol 'ʔ' [29]. For example, *loʔʔo* ('good') written using a double apostrophe, such as *lo"o*, *beʔaa* ('see'), can be written as *be'aa*.

4. Dataset

We utilized parallel datasets for Wolaytta–English from the Tonja et al. [19] study and collected a monolingual Wolaytta dataset from several sources, such as Wolaytta Fana Radio, Wolaytta Wogeta Radio, and Wolaytta text books. Table 2 depicts the parallel and monolingual datasets used in this study and their distribution in terms of the number of parallel and monolingual sentences, types, tokens, and vocabulary size.

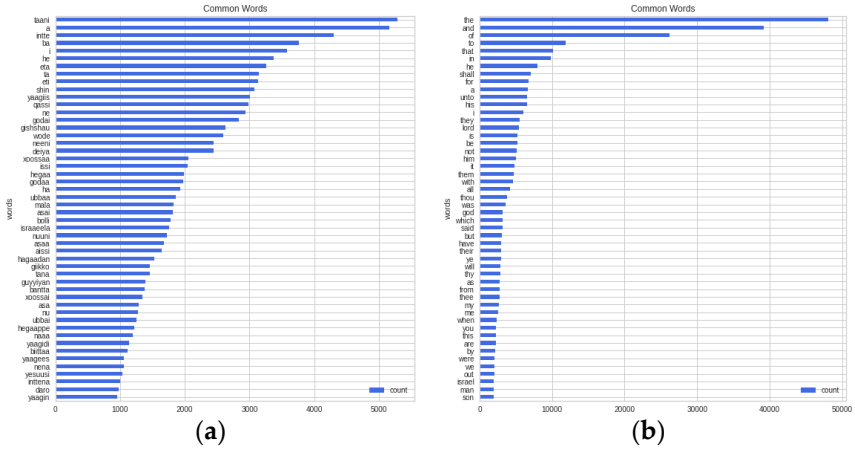
Table 2. Wolaytta–English parallel and Wolaytta monolingual dataset distribution.

Languages	Sentences	Tokens	Types	Vocabulary Size
English	30,495	763,083	19,736	11,130
Wolaytta		511,426	54,825	22,706
Wolaytta monolingual	40 k	496,251	70,289	

As Table 2 shows, our final dataset contains 30 k parallel sentences for the Wolaytta–English language pair and 40 k monolingual dataset for Wolaytta. In the Wolaytta–English parallel dataset, there are 19,736 unique words in English, 54,825 unique words in Wolaytta, and 70,289 unique words in the Wolaytta monolingual dataset.

Comparing unique words in the parallel dataset, Wolaytta has about three times more words than English. This is due to the morphological inflection and derivation of the Wolaytta language, which leads to a huge number of variations for a single word. Comparing the Wolaytta parallel dataset and the Wolaytta monolingual dataset, in the same way, the latter has about twice as many words as the parallel dataset.

Figures 1 and 2 show the most common tokens in the parallel and monolingual datasets, respectively. Parallel and monolingual Wolaytta data, as shown in Figures 1a and 2, share the most common words regardless of the domain. The Wolaytta–English parallel dataset belongs to the religion domain, but the Wolaytta monolingual data belong to the sports, news, and education domains. For example, words such as, *a* ('she'), *ha* ('this'), *ba* ('go'), *i* ('he'), *he* ('that'), *eta* ('them'), *ta* ('me'), *ne* ('you'), *gishawi* ('because of'), *deiya* ('have'), *issi* ('one'), and *mala* ('similar') appear in both datasets, but words such as *Xoossaa* ('God'), *israaeela* (Israel'), and *godai* ('The God') are not found in the Wolaytta monolingual data because these words are only used in the religion domain.



(**a**) (**b**)

Figure 1. Fifty most common words in the parallel corpora. (**a**) Wolaytta parallel data; (**b**) English parallel data.

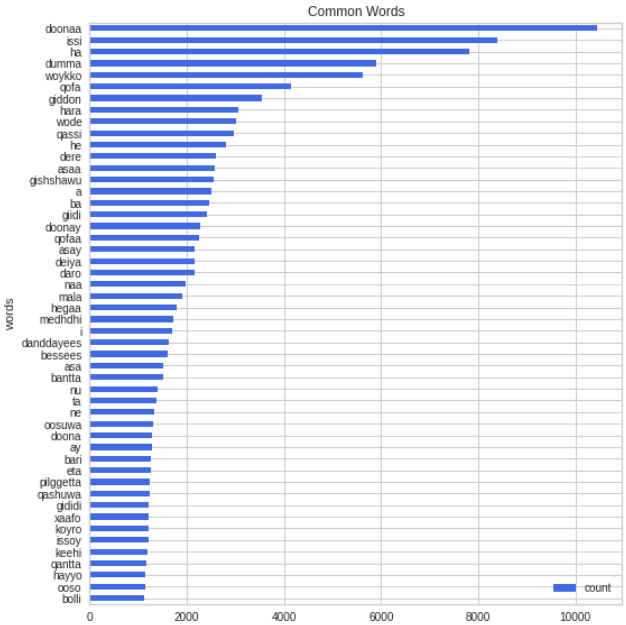
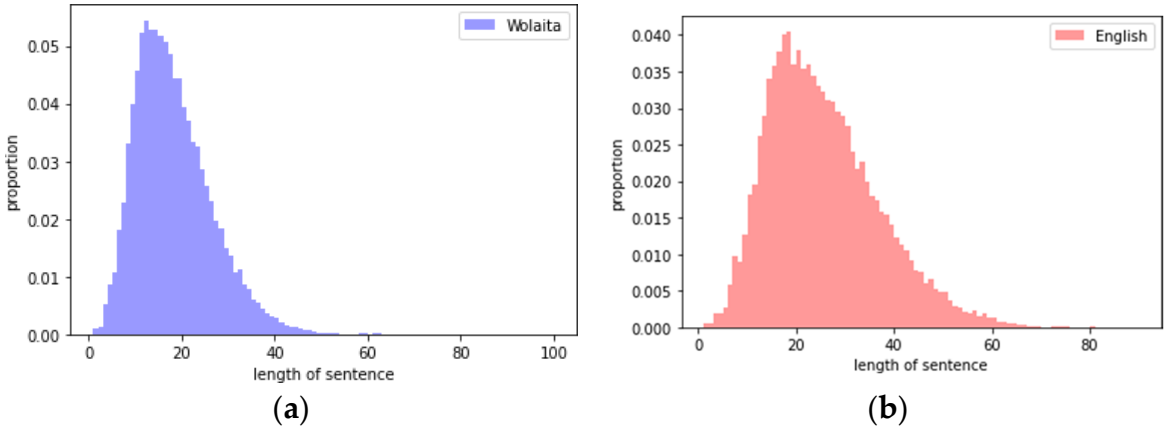


Figure 2. Fifty most common words in Wolaytta monolingual data.

Figure 3 depicts the word count per sentence in the parallel datasets for each of the languages. From the combined diagram of the Wolaytta and English parallel data in Figure 4, it is evident how different Wolaytta and English sentence lengths are.



(**a**) (**b**)

Figure 3. Sentence word count in (**a**) Wolaytta and (**b**) English parallel datasets. As can be observed in (**a**,**b**), the word count per sentence in Wolaytta and English is within the range [2–40].

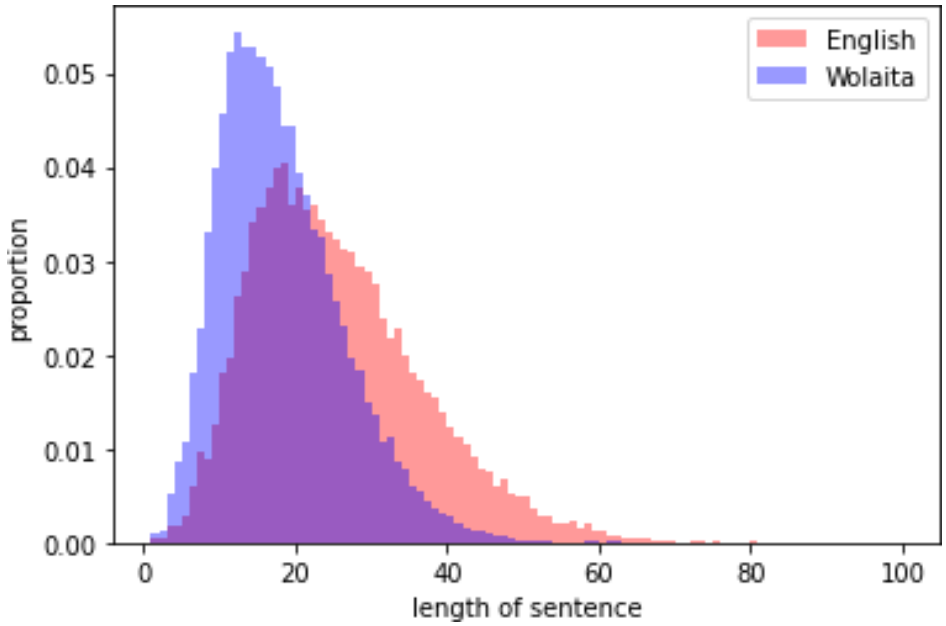


Figure 4. Sentence word count in Wolaytta–English parallel dataset.

Figures 5 and 6 include some text samples from the parallel and monolingual datasets. Table 3 Manually translated sample text for Figure 6 monolingual text shows translated sample text for the Wolaytta monolingual dataset shown in Figure 6.

index	Wol	Eng
0	a na'ai bairai abddoona a naati harati xuura qiisa ba'aala neera nadaaba	and his firstborn son abdon then zur and kish and baal and ner and nadab
1	awai mokkiyo baggaappe sooriyaa biittaa asatinne awai wulliyo baggaappe pilisxxeema biittaa asati bantta doonaa hanggidi israa'eela miidosona he ubbankka godaa hanqqoi simmibeenna a kusheekka biron denddidosan de'ees	the syrians before and the philistines behind and they shall devour israel with open mouth for all this his anger is not turned away but his hand is stretched out still
2	qassi issoi issuwaa go''iyaabaa koyanaappe attin intte huuphe xalaalaa go''iyaabaa koyoppite	look not every man on his own things but every man also on the things of others
3	'ha oiddu gita do'ati ha sa'an sinttappe denddana oiddu kawotettata	these great beasts which are four are four kings which shall arise out of the earth
4	iita ooso ubbaappe haakkite	abstain from all appearance of evil
5	un''o penggiyaara gelite aissi giikko bashshau efiya penggee aaho ogeekka woggaa hegaara geliya asaikka daro	enter ye in at the strait gate for wide is the gate and broad is the way that leadeth to destruction and many there be which go in thereat
6	shin kiitettidaageeti hayqqi simmin kaalliyaageetakka bantta geeduwaa kaalettanau worddduwaa haasayiya asati denddidosona	however especially after the death of the apostles men arose who spoke twisted things in order to draw away the disciples after themselves
7	yaanikko nena yihoowawu geppada xammaqettiyoogee bessiyaaba	then it would be fitting to dedicate your life to jehovah and get baptized
8	yaagin yesuusi zaaridi hagaadan yaagiis higgiyaa tamaarissiyaageetoo inttessi aayye'ana aissi giikko tookkanau deexxiya toohuwaa asa toosseeta shin tookkiya asa maaddanau harai atto intte intte huuphen biradhdhiyankka he toohuwaa bochchekketa	and he said woe unto you also ye lawyers for ye lade men with burdens grievous to be borne and ye yourselves touch not the burdens with one of your fingers
9	eti qassikka xomppiyaayyo sawiya ixaanaayyoonne tishshaassi haniya wogaraa zaitiyaanne qimamiyaa ehiidosona	and spice and oil for the light and for the anointing oil and for the sweet incense

Figure 5. Sample texts from Wolaytta–English parallel dataset.

index	Wolaytta Monolingual data
0	hegaa gishshawu doonaynne dere asay issoy issuwa bolli naaquwa gattiidi gidiyoogaa gidishin ha naaqoy dichchawukka hayquwawukka gaaso giyoogaa
1	ba soo pittennaara ba bollotee soo ona gawusu
2	qofaa shishshiyoosaa
3	qassikka doonaa malaatateetta gidennan doonay aattyioo kiita
4	hegaa be'ida eeyya away soo biidaagee ba booza mchchees binne na'ay heemmanawu nuuni yeddin asa karetta sawo wuuqqi miiddi doonaa kunttidi zin"iis yaagidi ba machcheessi yootiis
5	dumma dumma wolayttatto xufeta
6	yafaraa malaatadan zaara
7	eraa xekkaa nabbabuwa wogaa bariya haasayiyo doonaa de'iuwa hanotaainjjetiyoogaanne injjetennaagaa tamaariyo kifiliyanne peeshshaa xeelliyagan kumetta naqaashaa haarida asa gidanwu bessees
8	buraariya giyo keettaa qommuwaa keexxanaassi daro metootennan heeran beettida godaa hoottaa pariyaa woyshshaa uusunttaanne maqqaa shiishshidi keexxana danddayyettees
9	pilggetta huuphe qofakka shakkiyooga

Figure 6. Sample texts from Wolaytta monolingual dataset.

Table 3. Manually translated sample text for monolingual text Figure 6.

Wolayta Monolingual Equivalent Translation (Manually Translated)
So that the language and people put a victim one another, which resulted in the growth and death of it
she is saying that go to home whom not cleaning her mother-in-law dirty home
Minute or idea collection
Not only the sign of language but also the message it conveys
the rude father who saw it went to his home and said to his careless wife that our son we let him to keep cattle who stole black piece of powder and laid filling his mouth
different kinds of Wolaytta literature
answer as indicated in the paragraph
it must be based on the knowledge level of language reading culture that may or may not possible

5. Methodology

In this section, we present our experimental pipeline. The first step is data preprocessing, which is explained in Section 5.1. In Section 5.2, we discuss the hyper parameters used in the experiments. In Section 5.4, we present the experimental architecture of our NMT model followed by its detailed discussion. In Section 5.4.1, we go over the three models we trained on the parallel data to choose our baseline models. Then, we show how we used a source-side monolingual dataset and self-learning approach to improve the performance of our best baseline model. We explore self-training on the source-side monolingual (Wolaytta) dataset with the target-side synthetic (English) dataset and parallel dataset in Section 5.4.2, and in Section 5.4.3, we discuss the final experimental approach. Finally, in Section 5.5, we discuss the metrics used to evaluate our NMT models.

5.1. Data Preprocessing

Before training the NMT models with the datasets described in Section 4, we preprocessed both monolingual and parallel data in the following steps: (i) removing the duplicate sentences from the parallel dataset, (ii) converting all texts to lower case, (iii) removing special characters from texts except the apostrophe symbol for the reason discussed in Section 4 (Wolaytta uses the apostrophe (single quote) symbol to represent the glottal stop symbol (ʔ), (iv) tokenizing the source and target parallel sentences into subword tokens using Byte Pair Encoding (BPE) [33] representation, and (v) generating subword embeddings as input to the positional encoder block of the transformer model.

5.2. Hyperparameters

For our experiments, we selected hyperparameters according to the models we used. For the long short-term memory (LSTM) baseline model, as described in [34], we used an embedding layer with an embedding vector size of 512, layers of LSTMs (encoder and decoder) with 1000 neurons, and a dense layer with 1000 neurons with the Softmax activation in the output layer and a dropout of 0.2. We trained the model in 25 k steps with a batch size of 64. For a bidirectional long short-term memory (Bi-LSTM) model, we used a similar embedding size of 512, layers of LSTMs (encoder and decoder) with 512 neurons,

and a dense layer with 1024 neurons with the Softmax activation in the output layer and a dropout of 0.3. We trained the model in 25 k steps with a batch size of 64.

We chose the state-of-the-art transformer network proposed in [35], which consists of an encoder with six layers and a decoder with six layers. We used a transformer-big configuration from [36] in all experiments: the dimensions of the word embedding and the inner feed-forward layers are 512 and 2048, respectively. The number of attention heads is 16, with a dropout of 0.2. We trained the model in 25k steps with a batch size of 3072.

All models were trained using the Adam optimizer [37] with 1 = 0.9 and 2 = 0.98 and the cross-entropy loss. The padding was masked when computing the loss. The transformer used the learning rate scheduler defined in the original paper 'Attention is all you need' [35], while the LSTM and Bi-LSTM used a learning rate of 0.0001. The models were saved at the steps that showed the best validation bilingual evaluation understudy (BLEU) score.

5.3. Experimental Setting and Setup

We used the same environment as in [36] and trained all models in Google Colab Pro + [38] with OpenNMT [39]. We employed BPE [33] subword tokenization, and the BPE representation was chosen in order to remove vocabulary overlap during dataset combinations.

5.4. Experimental Architecture

Our experimental architecture is composed of three models, namely the baseline, self-trained, and final NMT models as shown in Figure 7. Experimental architecture of the Wolaytta–English NMT model. The detailed description of the models is discussed in Sections 5.4.1 and 5.4.2. To train the baseline model, we utilized the available Wolaytta–English parallel dataset [19], and to train the self-trained model, we used a combination of the authentic parallel sentences, Wolaytta monolingual, and English synthetic datasets. Finally, we fine-tuned the self-trained NMT model on the authentic parallel dataset to obtain the final NMT model.

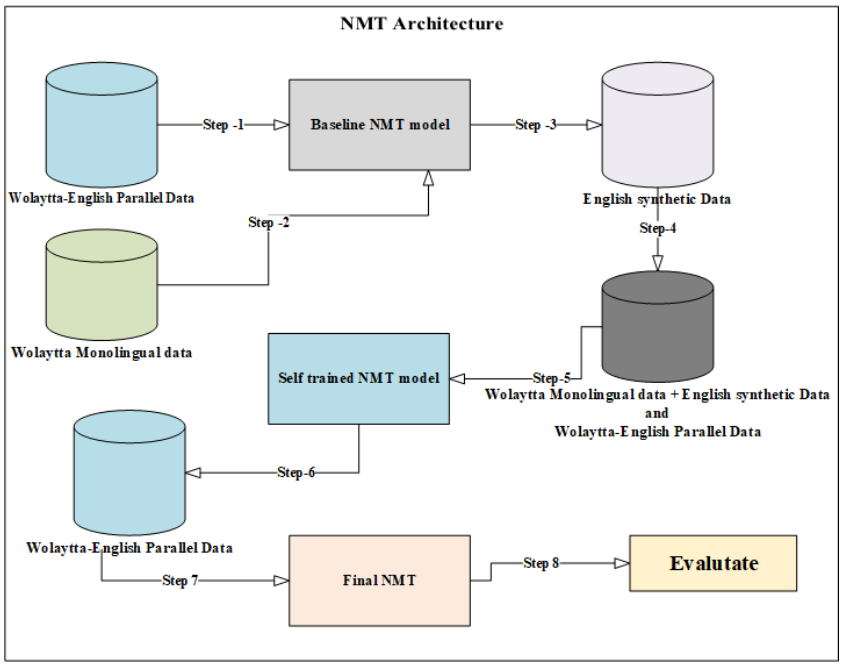


Figure 7. Experimental architecture of the Wolaytta–English NMT model.

The procedure consists of eight steps (described in Sections 5.4.1 and 5.4.2). Step 1: use the parallel dataset, train the baseline model; Steps 2 and 3: generate synthetic English data; Step 4: combine synthetic and authentic parallel datasets; Step 5: train the self-trained NMT

model in mixed data; Steps 6 and 7: fine-tune self-trained NMT model on the authentic parallel dataset; Step 8: evaluate the final NMT model on test data.

5.4.1. Baseline

To select the best-preforming model for further experiments, we conducted three baseline experiments using the authentic Wolaytta–English parallel dataset as depicted in Step 1 in Figure 7. Experimental architecture of the Wolaytta–English NMT model. For our baseline experiments, we used bidirectional LSTM encoder–decoders, transformer models, and unidirectional LSTM encoder–decoders. The goal of this experiment was to identify and select the model that outperformed on the authentic parallel dataset and then use the selected model in the rest of the experiments.

- LSTM is a special type of recurrent neural network introduced to solve the problem of vanishing and exploding gradients. This network also works better at keeping long-distance connections and figuring out how the values at the beginning and end of a sequence are related. LSTM only preserves information because the only input to an LSTM unit has seen is from the past that is from previous units.
- Bi-LSTM is a recurrent neural network used primarily for natural language processing (NLP). In contrast to standard LSTM, the input flows both ways, so the model can use information from both sides of a sequence. So, it is a very useful tool for modeling how words and phrases depend on each other in both directions of the sequence.
- Transformer is a type of artificial neural network architecture designed to solve the problem of transforming input sequences into output sequences in deep learning applications. It was proposed to take advantage of attention [35] and repetition to handle dependencies between input and output data. This transduction model employs self-attention to compute representations of its inputs and outputs, unlike recurrent neural networks (RNNs), which use sequences. Due to this feature, the transformer allows for much more parallelization than RNNs during training.

We used hyperparameters discussed in Section 5.2 for each baseline model. The baseline models were trained using the parallel Wolaytta–English parallel dataset. From the three baseline models, we then selected the best-preforming model to generate a synthetic English dataset with the help of the Wolaytta monolingual dataset. We combined the Wolaytta–English parallel dataset with the synthetic parallel dataset in the following training steps as described in Section 5.4.2.

5.4.2. Self-Learning

In machine translation, a model is trained on parallel (authentic) data and then used to translate a set of monolingual source sentences into the target language. This creates pseudo-parallel (synthetic) training data. In the next step, the synthetic data are used to train a better model by combining the pseudo-parallel data with the authentic parallel data. The self-training approach was first introduced by Ueffing et al. [18] to improve phrase-based statistical machine translation systems. In their work, they used a self-training system to generate translations of a set of source data. Then, the confidence score of each of the translated sentences was calculated, and based on these scores, reliable translations were selected and further adopted as additional training data for improving the same system.

We applied the self-learning approach to train the best-preforming model from the baseline models by combining the pseudo-parallel training data with the authentic parallel data. In our scenario, the self-learning approach was implemented in the following ways (more details can be found in Algorithm 1): in order to select the best-preforming model from our baseline models, we trained three baseline NMT models Tm for languages X and Y (X and Y represent Wolaytta and English, respectively) in the $X \rightarrow Y$ translation direction using the authentic X–Y parallel dataset Pd. After training Tm, we generated the target synthetic dataset Sd by translating the Wolaytta monolingual dataset Md using the selected Tm from our baseline model. Then, we combined Md and Sd to form the synthetic

parallel dataset Sp. Finally, we trained a new $X \rightarrow Y$ translation on the combination of both Pd and Sp datasets.

Algorithm 1 Self-learning approach.

Requires:
- Authentic parallel dataset: ***Pd***
- Monolingual dataset: ***Md***
- Target synthetic dataset: ***Sd***
- Synthetic parallel dataset: ***Sp***
- Languages: ***X***, ***Y***
- Translation model: ***Tm*** $\boldsymbol{X \rightarrow Y}$

Ensures:
Train ***Tm*** $\boldsymbol{: X \rightarrow Y}$ on ***Pd***
Generate ***Sd*** by translating ***Md*** using trained ***Tm*** $\boldsymbol{: X \rightarrow Y}$
Combine ***Md*** and ***Sd*** to form ***Sp***
Train final ***Tm*** $\boldsymbol{: X \rightarrow Y}$ on the combination of ***Pd*** and ***Sp***

In our self-learning approach, we used in-domain, mixed validation sets to fine-tune the parameters of the NMT model during training time, and evaluated its performance using an in-domain test set.

5.4.3. Fine-Tuning

Transfer learning is one of the currently suggested approaches to improve the performance of NMT for low-resource languages [12]. We applied transfer learning by fine-tuning our self-trained NMT models using the parallel Wolaytta–English parallel dataset to obtain our final NMT model. During fine-tuning, we used both self-trained NMT models with in-domain and mixed validation sets.

5.5. Evaluation

We evaluated the performance of our model in terms of translation accuracy. There are many evaluation techniques developed for such purpose: human evaluation, Bi-lingual Evaluation Understudy (BLEU) score, National Institute of Standards and Technology (NIST) score, Metric for Evaluation of Translation with Explicit Ordering (METEOR), Translation Edit Rate (TER), and Character-level F-score (chrf).

5.5.1. BLEU Score

BLEU is an automatic metric based on n-grams [40]. It measures MT adequacy by looking at word precision and MT fluency by calculating n-gram precisions, returning a translation score with in the range [0, 1] or alternatively with a [0, 100] scale. We used a particular implementation of BLEU, called SacreBLEU. It outputs dataset scores, not segment scores. The greater the score, the closer the translation is to the reference. The rationale behind BLEU is that high-quality translations will share many n-grams with human translations [40].

BLEU is defined as

$$BLEU = BP4 \times \left(\prod_{n=1}^{4} \times pn \right)$$

where pn measures the modified n-gram precision between a document with candidate translations and a set of human-authored reference documents, and the brevity penalty BP downscales the score for outputs shorter than the reference. *BLEU* measures the closeness of the machine translation to the human reference translation taking into consideration the translation length, word choice, and word order.

5.5.2. Translation Edit Rate (TER)

TER is a character-based automatic metric for measuring the number of edit operations needed to transform the machine-translated output into a human-translated reference [41]. TER is defined as the minimum number of edits needed to change a hypothesis (i.e., translation) so that it exactly matches one of the references, normalized by the average length of the references. Because we were concerned with the minimum number of edits needed to modify the hypothesis, we measured only the number of edits to the closest reference specifically [41]:

$$TER = \frac{\#edit\ average}{\#of\ referencewords}.$$

Possible edits include insertion, deletion, and substitution of single words as well as shifts of word sequences. The dataset TER score is the total number of edits divided by the total number of words and multiplied by 100. TER ranges from 0 to infinity. The greater the score, the farther the translation is from the reference.

5.5.3. Character-Level F-Score (chrF)

chrF is a tool for the automatic evaluation of machine translation output based on character n-gram precision and recall, enhanced with word n-grams [42]. It calculates the F-score averaged on all the character and word n-grams, with character n-gram order set to 6 and word n-gram order set to 2. chrF score calculated as

$$(1 + \beta 2)\frac{CHRP \times CHRR}{\beta 2 \times CHRP + CHRR},$$

where *CHRP* is the percentage of n-grams in the hypothesis, which have a counterpart in the reference (character n-gram precision), and *CHRR* is the percentage of character n-grams in the reference, which is also present in the hypotheses (character n-gram recall). *CHRP* and *CHRR* are averaged over all n-grams, and β is the parameter to assign β times more importance to recall than to precision. If β = 1, the precision and recall have the same importance.

6. Results

In this section, we discuss the results of our experiments. In Section 6.1, we show the results of our baseline NMT models in both Wolaytta–English and English–Wolaytta translations. In Section 6.2, we present the results of the self-trained NMT models using both synthetic and authentic datasets. Finally, in Section 6.3, we give the results of our final fine-tuned NMT model on the authentic dataset.

6.1. Baseline

As discussed in Section 5.4.1, to select the best-performing model, we trained and evaluated three baseline models on the authentic parallel dataset. Table 4 shows the BLEU, chrF, and TER scores of the three baseline models in Wolaytta-English translation. It can be seen in Table 4 that the best-performing baseline model is the transformer model. Similarly, Table 5 shows that for the English-Wolaytta translation, the transformer model outperformed the other models. In Section 7, we discuss the performance of the baseline models in detail.

Table 4. BLEU, chrF, and TER scores of baseline models in Wolaytta–English translation.

Model (Wolaytta–English)	BLEU (%)	chrF	TER	Loss
LSTM	6.0	23.9	87.4	5.38
Bi-LSTM	6.7	24.7	86.6	**5.30**
Transformer	**12.2**	**31.8**	**80.2**	5.69

Table 5. BLEU, chrF, and TER scores of baseline models in English–Wolaytta translation.

Model (English–Wolaytta)	BLEU (%)	chrF	TER	Loss
LSTM	2.1	18.2	94.9	**6.18**
Bi-LSTM	2.8	20.5	95.0	7.28
Transformer	**6.2**	**26.0**	**93.9**	6.26

6.2. Self-Learning

As considered in Section 5.4.2, we selected the outperforming NMT model from our baseline models and trained it on the combination of mixed (parallel and synthetic) dataset. Tables 6 and 7 show the BLEU, chrF, and TER scores of the self-trained NMT models trained on a combination of the authentic and synthetic parallel datasets for Wolaytta–English and English–Wolaytta translation, respectively.

Table 6. BLEU, chrF, and TER scores of NMT models trained on the combination of synthetic and authentic datasets for Wolaytta–English translation.

Model (English–Wolaytta)	BLEU (%)	chrF	TER	Loss
Transformer (mixed validation set)	14.7	34.7	78.1	3.72
Transformer (in-domain validation)	**14.9**	**35.1**	**77.6**	**3.61**

Table 7. BLEU, chrF, and TER scores of NMT models trained on the combination of synthetic and authentic datasets for English–Wolaytta translation.

Model (English–Wolaytta)	BLEU (%)	chrF	TER	Loss
Transformer (mixed validation set)	8.4	30.3	89.7	**6.62**
Transformer (in-domain validation set)	**8.6**	**30.6**	**88.9**	6.98

6.3. Fine-Tuning

As shown in Section 5.4.3, to train the final NMT model, we fine-tuned the self-trained NMT models in both in-domain and mixed validation sets using the authentic parallel datasets. Tables 8 and 9 present the results of fine-tuning on the authentic parallel dataset for Wolaytta–English and English–Wolaytta translations, respectively.

Table 8. BLEU, chrF, and TER scores of NMT models fine-tuned on the authentic dataset for Wolaytta–English translation.

Model (English–Wolaytta)	BLEU (%)	chrF	TER	Loss
Fine-tuned (mixed validation set)	15.7	35.8	76.5	4.13
Fine-tuned (in-domain validation set)	**16.1**	**36.0**	**74.5**	**3.53**

Table 9. BLEU, chrF, and TER scores of NMT models fine-tuned in authentic dataset for English–Wolaytta translation.

Model (English–Wolaytta)	BLEU (%)	chrF	TER	Loss
Fine-tuned (mixed validation set)	8.7	30.4	88.6	4.82
Fine-tuned (in-domain validation set)	**9.0**	**31.8**	**86.1**	**4.51**

7. Discussion

In this section, we discuss the performance of our NMT models. In Section 7.1, we consider the performance of the three baseline NMT models followed by a discussion of the performance of the self-trained NMT models in Section 7.2. Finally, in Section 7.3, we discuss the performance of our final NMT model. To evaluate the performance of our NMT models, we used BLEU, chrF, and TER automatic evaluation metrics. Higher BLEU and

chrF scores mean that a the translation is closer to the reference (test set), while a higher TER score indicates that a the translation is farther from the reference (test set).

7.1. Baseline

Tables 4 and 5 show the results of our baseline experiments for Wolaytta–English and English–Wolaytta translations, respectively. For Wolaytta–English translation, the Uni-LSTM and Bi-LSTM models demonstrated 6.0 and 6.7 BLEU; 23.9 and 24.7 chrF; and 87.4 and 86.6 TER scores, respectively. Based on these results, we can confirm that the encoder–decoder Bi-LSTM model's performance is better than that of the encoder–decoder Uni-LSTM model, which is explained by the additional layer aggregation in Bi-LSTM: this enabled the model to perform better than the Uni-LSTM. Similarly, for English–Wolaytta translation, the Uni-LSTM and Bi-LSTM models showed 2.1 and 2.8 BLEU; 18.2 and 20.5 chrF; and 97.0 and 95.0 TER scores, respectively. The numbers make it evident that the Bi-LSTM model's performance is better than that of Uni-LSTM; the former took advantage of the additional layer to predict a better reference than Uni-LSTM. Comparing all the results in both encoder–decoder models, we see that the translation into English as the target language was of better quality than the translation from English as the source language; this is due to the morphological complexity of the Wolaytta language and the technological favor of the model for a high-resource language.

Comparing the transformer model with both encoder–decoder models, the Transformer outperformed both models in both translations. For Wolaytta–English translation, the transformer model showed a 12.2 BLEU score, which is two times greater than that of the Bi-LSTM model; this shows that the input sequence processing technique and the attention mechanism of the Transformer boosted the performance of the model compared to the Uni-LSTM and Bi-LSTM models. In the latter models, we did not use an attention mechanism, and the way they process the input sequence is different from that of the Transformer. The Uni-LSTM and Bi-LSTM models process sequences of symbol representations step by step and separately, while the Transformer processes all the input sequences in one step. Similarly for English–Wolaytta translation, the Transformer model showed a 6.2 BLEU score, which is three times higher than that of Bi-LSTM. Observing the performance of the transformer in both translations, we see that it is challenged when English is the source language compared to when English is the target language.

Thus, from the three baseline models, the Transformer model outperformed the other models; therefore, we selected the Transformer model for the rest of the experiments. As we have already discussed, the goal of the baseline experiment was to choose the model that works best for our experiments.

7.2. Self-Learning

Tables 6 and 7 present the results of the self-learning experiments for Wolaytta–English and English–Wolaytta translations, respectively. We conducted two experiments to investigate the impact of using source-side monolingual data to improve NMT for a low-resource language, as discussed in Section 5.4.2. In the experiments, we combined synthetic and authentic parallel datasets using two methods. In the first experiment, we used a mixed validation set obtained by combining synthetic and authentic parallel datasets further splitting the resulting dataset into training and validation sets in the ratio of 80:20, respectively, and then we applied a test set from the authentic parallel dataset. In the second experiment, after training the model with both synthetic and authentic parallel datasets, we used validation and test sets from the same domain.

As we can see from the results, training the selected transformer model on a combination of synthetic and authentic parallel datasets showed an increment in the BLEU, chrF, and TER scores of the NMT model in both translation directions. For Wolaytta–English translation, applying the self-learning approach in the mixed validation set and the in-domain validation set showed 14.7 and 14.9 BLEU; 34.7 and 35.1 chrF; and 78.1 and 77.6 TER scores, respectively. As it can be seen from the results, the model trained on a combination of

synthetic and authentic parallel datasets and validated on an in-domain set outperformed the model trained and validated on a mixed dataset by +0.2 BLEU, +0.4 chrF, and −0.5 TER scores. This shows that using an in-domain validation set helped the NMT model fine-tune hyperparameters during the training phase, which led to better performance than using a mixed validation set. We observed that mixing synthetic and authentic parallel datasets improved NMT quality by the +2.5–2.7 BLEU, +2.9–3.3 chrF, and −2.6–2.1 TER scores for the Wolaytta–English translation compared with the baseline transformer model.

In the same way, applying the self-learning method to the mixed validation set and the in-domain validation set produced 8.4 and 8.6 BLEU; 30.3 and 30.6 chrF; and 89.7 and 88.9 TER scores for English–Wolaytta translation, respectively. For the English–Wolaytta translation, the model trained on the in-domain validation set outperformed the model trained on the mixed validation set. When the overall performance of the English–Wolaytta translation was compared to the baseline transformer model, it was found that mixing synthetic and authentic parallel datasets improved NMT quality by +2.2–2.4 BLEU, +4.3–4.6 chrF, and −4.2–5 TER scores.

Therefore, based on the above results, we can say that utilizing only source-side monolingual data improves the performance of NMT for Wolaytta as a low-resource language in both translation directions if we combine synthetic and authentic parallel datasets for training and use authentic datasets for validation and testing.

7.3. Fine-Tuning

We fine-tuned each selected model for up to 10k steps using early stopping on the validation set, and we set the validation score threshold to stop training at BLEU score of 0.2 over the last four iterations. By using the authentic parallel datasets, we fine-tuned the two models described in Section 7.2 to further improve the NMT performance of Wolaytta–English translation in both directions. The results of the fine-tuning self-trained NMT models for Wolaytta–English translation using mixed validation and in-domain validation sets show 15.7 and 16.1 BLEU; 35.8 and 36.0 chrF; and 76.5 and 74.5 TER scores, respectively. Fine-tuning the NMT models trained on authentic parallel datasets increased the NMT quality by +0.8–1.2 BLEU, +0.7–0.9 chrF, and −1.1–3.1 TER scores compared to the NMT models trained on a mix of authentic and synthetic parallel datasets.

In the same way, fine-tuning of the self-trained NMT models for English–Wolaytta translation using mixed validation and in-domain validation sets led to 8.7 and 9.0 BLEU; 30.4 and 31.8 chrF; and 88.6 and 86.1 TER scores, respectively. This showed the improvement of NMT quality by +0.3–0.6 BLEU, +0.2–1.2 chrF, and −0.9–2.8 TER scores over the top-performing NMT model trained on the mixed dataset.

Based on the above results, we can say that using the authentic parallel dataset to fine-tune NMT models trained on both synthetic and authentic data improves the NMT performance in both directions for low-resource Wolaytta–English. The results of the test sets show that fine-tuning the NMT models that were trained on in-domain validation sets improves performance in both directions of translation.

7.4. Limitation of the Study

In this paper, we did not evaluate the confidence score, which is widely used in the evaluation of translation quality. Further study of the translation qualities of monolingual data sets before combining them with authentic datasets during the training phase would improve the NMT performance only by sampling the dataset based on the confidence score. In addition, the NMT system for low-resource languages would be better if it was possible to use a monolingual dataset of a low-resource language in a multilingual setting.

8. Conclusions

In this paper, we studied whether a source-side monolingual dataset of a low-resource language could be used to improve the NMT system for such language. As a low-resource language translation example, we used the Wolaytta-English language pair. We used an

approach called self-learning and fine-tuning, along with synthetic and authentic parallel datasets for the Wolaytta-English language pair in both directions of translation. We showed that combining synthetic and authentic parallel datasets of a low-resource language in a self-learning method led to improvements in the BLEU scores of +2.5–2.7 and +2.2–2.4 for Wolaytta-English and English-Wolaytta translations, respectively, over the best-performing baseline. In addition we showed that fine-tuning the NMT models trained in a self-learning approach on an authentic parallel Wolaytta-English dataset improved the BLEU score over the self-learning approach by +0.8–1.2 and +0.3–0.6 for Wolaytta-English and English-Wolaytta translations, respectively.

In the future, we would like to study the effect of increasing the size of the source-side monolingual data in a low-resource NMT system. We will also investigate the benefits of using source-side monolingual data for languages similar to Wolaytta that do not have a monolingual dataset, and we will investigate if the proposed approach works for other low-resource languages.

Author Contributions: Conceptualization, A.L.T. and O.K.; Methodology, A.L.T., O.K., A.G. and G.S.; Software, G.S.; Validation, A.G.; Formal analysis, A.L.T. and O.K.; Investigation, A.L.T., O.K., A.G. and G.S.; Resources, A.L.T. and O.K.; Data curation, A.L.T.; Writing—original draft, A.L.T.; Writing—review & editing, O.K. and G.S.; Visualization, A.G.; Supervision, O.K. and A.G.; Project administration, A.L.T., O.K. and G.S.; Funding acquisition, O.K., A.G. and G.S. All authors have read and agreed to the published version of the manuscript.

Funding: The work was performed with partial support from the Mexican Government through the grant A1S-47854 of CONACYT, Mexico and 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset and the script used in this paper can be accessed by https://github.com/atnafuatx/EthioNMT-datasets.

Conflicts of Interest: The authors declare no conflict of interest.

Glossary

Authentic dataset	original parallel dataset.
Source side	language whose text is to be translated.
Synthetic dataset	pseudo-data generated using monolingual data.
Target side	language into which the source text is to be translated.

References

1. Madankar, M.; Chandak, M.B.; Chavhan, N. Information retrieval system and machine translation: A review. *Procedia Comput. Sci.* **2016**, *78*, 845–850. [CrossRef]
2. Kenny, D. Machine translation. In *The Routledge Handbook of Translation and Philosophy*; Routledge: Oxfordshire, UK, 2018; pp. 428–445.
3. Bahdanau, D.; Cho, K.H.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
4. Forcada, M.L. Making sense of neural machine translation. *Transl. Spaces* **2017**, *6*, 291–309. [CrossRef]
5. Nekoto, W.; Marivate, V.; Matsila, T.; Fasubaa, T.; Kolawole, T.; Fagbohungbe, T.; Akinola, S.O.; Muhammad, S.H.; Kabongo, S.; Osei, S.; et al. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv* **2020**, arXiv:2010.02353.
6. Freitag, M.; Firat, O. Complete multilingual neural machine translation. *arXiv* **2020**, arXiv:2010.10239.
7. Ahmadnia, B.; Dorr, B.J. Augmenting neural machine translation through round-trip training approach. *Open Comput. Sci.* **2019**, *9*, 268–278. [CrossRef]
8. Johnson, M.; Schuster, M.; Le, Q.V.; Krikun, M.; Wu, Y.; Chen, Z.; Thorat, N.; Viégas, F.; Wattenberg, M.; Corrado, G.; et al. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 339–351. [CrossRef]

9. Aharoni, R.; Johnson, M.; Firat, O. Massively multilingual neural machine translation. *arXiv* **2019**, arXiv:1903.00089.
10. Koehn, P.; Knowles, R. Six challenges for neural machine translation. *arXiv* **2017**, arXiv:1706.03872.
11. Lakew, S.M.; Federico, M.; Negri, M.; Turchi, M. Multilingual neural machine translation for low-resource languages. *Ital. J. Comput.* **2018**, *4*, 11–25. [CrossRef]
12. Zoph, B.; Yuret, D.; May, J.; Knight, K. Transfer learning for low-resource neural machine translation. *arXiv* **2016**, arXiv:1604.02201.
13. Goyal, V.; Kumar, S.; Sharma, D.M. Efficient neural machine translation for low-resource languages via exploiting related languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Online, 5–10 July 2020; pp. 162–168.
14. Chowdhury, K.D.; Hasanuzzaman, M.; Liu, Q. Multimodal neural machine translation for low-resource language pairs using synthetic data. In Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP, Melbourne, Australia, 19 July 2018; pp. 33–42.
15. Fadaee, M.; Bisazza, A.; Monz, C. Data augmentation for low-resource neural machine translation. *arXiv* **2017**, arXiv:1705.00440.
16. Imankulova, A.; Sato, T.; Komachi, M. Improving low-resource neural machine translation with filtered pseudo-parallel corpus. In Proceedings of the 4th Workshop on Asian Translation (WAT2017), Taipei, Taiwan, 27 November–1 December 2017; pp. 70–78.
17. Gu, J.; Wang, Y.; Chen, Y.; Cho, K.; Li, V.O.K. Meta-learning for low-resource neural machine translation. *arXiv* **2018**, arXiv:1808.08437.
18. Ueffing, N.; Haffari, G.; Sarkar, A. Transductive learning for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 25–27 June 2007; pp. 25–32.
19. Tonja, A.L.; Woldeyohannis, M.M.; Yigezu, M.G. A Parallel Corpora for bi-directional Neural Machine Translation for Low Resourced Ethiopian Languages. In Proceedings of the 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA), Bahir Dar, Ethiopia, 22–24 November 2021; pp. 71–76.
20. Laskar, S.R.; Paul, B.; Adhikary, P.K.; Pakray, P.; Bandyopadhyay, S. Neural Machine Translation for Tamil–Telugu Pair. In Proceedings of the Sixth Conference on Machine Translation, Online, 10–11 November 2021; pp. 284–287.
21. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
22. Marie, B.; Fujita, A. Synthesizing Monolingual Data for Neural Machine Translation. *arXiv* **2021**, arXiv:2101.12462.
23. Tars, M.; Tättar, A.; Fišel, M. Extremely low-resource machine translation for closely related languages. *arXiv* **2021**, arXiv:2105.13065.
24. Sennrich, R.; Haddow, B.; Birch, A. Improving neural machine translation models with monolingual data. *arXiv* **2015**, arXiv:1511.06709.
25. Jiao, W.; Wang, X.; Tu, Z.; Shi, S.; Lyu, M.R.; King, I. Self-training sampling with monolingual data uncertainty for neural machine translation. *arXiv* **2021**, arXiv:2106.00941.
26. Dione, C.M.B.; Lo, A.; Nguer, E.M.; Ba, S. Low-resource Neural Machine Translation: Benchmarking State-of-the-art Transformer for Wolof<-> French. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 6654–6661.
27. Pham, N.L. Data Augmentation for English-Vietnamese Neural Machine Translation: An Empirical Study. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4216607 (accessed on 7 January 2023).
28. Ngo, T.-V.; Nguyen, P.-T.; Nguyen, V.V.; Ha, T.-L.; Nguyen, L.-M. An Efficient Method for Generating Synthetic Data for Low-Resource Machine Translation: An empirical study of Chinese, Japanese to Vietnamese Neural Machine Translation. *Appl. Artif. Intell.* **2022**, *36*, 2101755. [CrossRef]
29. Wakasa, M. A Descriptive Study of the Modern Wolaytta Language. Unpublished Ph.D. Thesis, University of Tokyo, Bunkyō, Tokyo, 2008.
30. Hirut, W. Writing both difference and similarity: Towards a more unifying and adequate orthography for the newly written languages of Ethiopia: The case of Wolaitta, Gamo, Gofa and Dawuro. *J. Lang. Cult.* **2014**, *5*, 44–53. [CrossRef]
31. Dalke, D. Tense, Aspect and Mood (TAM) in Wolayta. Ph.D. Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2012; p. 571.
32. Lessa, L. Development of Stemming Algorithm for Wolaytta Text. Ph.D. Thesis, Addis Ababa University, Addis Ababa, Ethiopia, 2003.
33. Gage, P. A new algorithm for data compression. *C Users J.* **1994**, *12*, 23–38.
34. Arif, M.; Tonja, A.L.; Ameer, I.; Kolesnikova, O.; Gelbukh, A.; Sidorov, G.; Meque, A.G.M. CIC at CheckThat! 2022: Multi-class and cross-lingual fake news detection. *Work. Notes CLEF* **2022**, *3180*, 434–443.
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
36. Tonja, A.L.; Kolesnikova, O.; Arif, M.; Gelbukh, A.; Sidorov, G. Improving Neural Machine Translation for Low Resource Languages Using Mixed Training: The Case of Ethiopian Languages. In Proceedings of the Advances in Computational Intelligence, Monterrey, Mexico, 24–29 October 2022; pp. 30–40.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Canesche, M.; Bragança, L.; Neto, O.P.V.; Nacif, J.A.; Ferreira, R. Google Colab CAD4U: Hands-On Cloud Laboratories for Digital Design. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 22-28 May 2021; pp. 1–5.
39. Klein, G.; Kim, Y.; Deng, Y.; Nguyen, V.; Senellart, J.; Rush, A.M. OpenNMT: Neural machine translation toolkit.*arXiv* **2018**, arXiv:1805.11462.

40. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Stroudsburg, PA, USA, 7–12 July 2002; pp. 311–318.
41. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
42. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal, 17–18 September 2015; pp. 392–395.