Addis Ababa University

Addis Ababa Institute of Technology

School of Information Technology and Engineering (SiTE)

MSc in Artificial Intelligence

Enhancing Neural Machine Translation Through Incorporation of Unsupervised Language Understanding and Generation Techniques: The Case of English-Afaan Oromo Translation

By

Chala Bekabil Geta

Supervisor

Fantahun Bogale (PhD)

A Thesis submitted to School of Information Technology and Engineering (SiTE) in partial fulfillment of the requirements for the degree of MSc in Artificial Intelligence

Submission Date: May 13, 2024

Addis Ababa, Ethiopia

May, 2024

Addis Ababa University
Addis Ababa Institute of Technology
School of Information Technology and Engineering (SiTE)

By: Chala Bekabil Geta

This is to certify that the thesis prepared by Chala Bekabil Geta, titled: Enhancing Neural Machine Translation Through Incorporation of Unsupervised Language Understanding and Generation Techniques: The Case of English-Afaan Oromo Translation, and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Artificial Intelligence complies with the regulations of the University and meets the accepted standards concerning originality and quality.

Approved by Board of Examiners

	Name	Signature	Date
Supervisor:	Fantahun B. (PhD)		11-JUN-2024
External Examiner:	Salomon T. (PhD)		11/06/2024
Internal Examiner:	Natnael Argaw		11/06/2024

Author's Declaration of Originality

I hereby certify that I am the sole author of this Thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Chala Bekabil Geta

06-06-2024

Acknowledgments

"Give thanks in all circumstances; for this is God's will for you in Christ Jesus." - 1 Thessalonians 5:18

I would like to express my deep appreciation and gratitude to the following individuals who have played a significant role in the completion of this thesis:

First and foremost, I am profoundly grateful to God for providing me with the wisdom, strength, and perseverance to undertake this journey. Your divine guidance has been a constant source of inspiration, and I am humbled by the blessings bestowed upon me.

I extend my heartfelt appreciation to my advisor, Fantahun B. (PhD), for his invaluable support and guidance. His expertise, patience, and unwavering commitment to my academic growth have been instrumental in shaping the direction of my research. I am truly fortunate to have had the opportunity to work under his supervision.

I am also grateful to my friends for their unwavering support and encouragement throughout this endeavor. Your belief in my abilities and willingness to assist whenever needed have been invaluable. I am appreciative of the countless hours we have spent discussing ideas, providing feedback, and uplifting each other.

To my family, I am eternally grateful for your unconditional love and unwavering support during this journey. Your belief in my capabilities and constant encouragement have been my driving force. I am indebted to you for your sacrifices, understanding, and unwavering belief in my potential.

I would like to extend my sincere gratitude to all those mentioned above, as well as to anyone who has supported me in various ways. Your unwavering support, guidance, and belief in me have been integral to the completion of this thesis. Thank you for being an essential part of my academic journey.

Abstract

Breaking down language barriers is a paramount pursuit in the realm of Artificial Intelligence. Machine Translation (MT), a domain within Natural Language Processing (NLP), holds the potential to bridge linguistic gaps and foster global communication. Enhancing cross-cultural communication through MT will be realized only if we succeed in developing accurate and adaptable techniques which in turn demands adequate availability of linguistic resources. Unluckily, under-resourced languages face challenges due to limited linguistic resources and sparse parallel data. Previous studies tried to solve this problem by using monolingual pre-training techniques. However, such studies solely rely on either Language Understanding (LU) or Language Generation (LG) techniques resulting in skewed translation. This study aims to enhance translation outcomes beyond the capabilities of previous studies by marrying the concepts of LU and LG and hence boosting the quality of MT in both directions. Our proposed model, the BERT-GPT incorporated Transformer, combines SOTA language models, BERT and GPT, trained on monolingual data into the original Transformer model and demonstrates substantial improvements. Experimental results shows that translation quality leaps forward, as evidenced by a significant increase in the BLEU score reaching 42.09, from the baseline score of 35.75 for English to Afaan Oromo translation, and 44.51 from the baseline score of 40.35 for Afaan Oromo to English translation on test dataset. Notably, our model unveils a deep understanding of Afaan Oromo's linguistic nuances, resulting in translations that are precise, contextually appropriate, and faithful to the original intent. By leveraging the power of unsupervised pre-training and incorporation of unsupervised LU and LG techniques to the transformer model, we pave the way for enhanced cross-cultural communication, advanced understanding and inclusivity in our interconnected world.

Keywords: BERT, GPT, NMT, Unsupervised pre-training

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Terms

ALPAC	Automatic Language Processing Advisory Committee
ANN	Artificial Neural Network
ASR	Automatic Speech Recognition
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BGRU	Bi-directional Gated Recurrent Unit
CBMT	Context-based Machine Translation
CNN	Convolutional Neural Network
CV	Computer Vision
GPT	Generative Pre-Training
GRU	Gated Recurrent Units
HTER	Human Target Error Rate
LSTM	Long Short Term Memory
LM	Language Model
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MT	Machine Translation
NIST	The National Institute of Standards and Technology
NL	Natural Language
NLP	Natural Language Processing
NMT	Neural Machine Translation
PTLM	Pre-training Language Model
SOTA	State Of The Art
RBMT	Rule-Based Machine Translation
RNN	Recurrent Neural Network
RNNMT	Recurrent Neural Network Machine Translation
SMT	Statistical Machine Translation
TER	Translation Error Rate

Chapter One

1. Introduction

1.1 Background

Natural Language (NL) is the method by which humans communicate with one another. This NL should be mutually comprehensible by the two communicating entities, which is improbable given that there are more than 7 billion people and around 7,139 officially known languages are spoken on the planet. Therefore, there should be a translator between the two, who is typically a human. However, using a human translation is costly and inconvenient. Natural Language Processing (NLP) is an area of Artificial Intelligence that focuses on enabling machines to understand, interpret, and generate natural languages [1]. One of the most common applications of NLP is Machine Translation (MT), which involves automatically translating text from one language to another. MT has the potential to break down language barriers and facilitate cross-cultural communication in various domains, including but not limited to foreign media, Global Conferences, Tourism, and International Trade [2, 3].

Although MT has made significant progress over the years, it still faces several challenges. One of the biggest challenges is the problem of generating accurate translations for under-resourced languages [4]. These are languages for which there is limited or no parallel data available for training MT models, as well as limited basic language resources such as dictionaries and grammars.

To address this problem, researchers have explored different approaches to MT. One approach is the use of Rule-based systems, which rely on predefined rules to translate text [5, 6, 7]. However, these systems are limited in their ability to handle complex sentence structures and language nuances. Another approach is Statistical Machine Translation (SMT), which involves training MT models on large amounts of parallel data [8, 9, 10, 11]. However, this approach requires significant amount of resources, and is not feasible for under-resourced languages.

More recently, the field of MT has seen a shift towards the use of Neural Machine Translation (NMT), which involves training MT models using Neural Networks (NN) [12, 13, 14]. This approach has shown promising results in improving the accuracy of translations, but still faces challenges in handling under-resourced languages. To address the challenges faced by under-resourced languages, researchers have explored the use of pre-training language models. Pre-training involves training a NN on a large amount of unlabeled data, and then fine-tuning the network on a specific task such as MT. Pre-trained language models have shown promising results to improve the performance of MT models for under-resourced languages. In NLP [15,

16, 17], Computer Vision (CV) [18, 19], and Automatic Speech Recognition (ASR) [20, 21, 22], pre-training is a prevalent paradigm. By leveraging the knowledge learned from pre-training, MT models can better handle the unique characteristics of under-resourced languages, even in the absence of parallel data and basic language resources. State of the art MT techniques previously tried to solve the current challenges in enhancing the problem of being faced by under resourced languages by focusing on unsupervised pre-training techniques. But they haven't explicitly studied deeply focusing on monolingual data with state-of-the-art NMT by balancing both understanding of source language and generating high quality translation for target language.

Therefore, this study aims to investigate the potential of unsupervised pre-training techniques in improving machine translation for English-Afaan Oromo, an under-resourced language for which there is limited parallel data and basic language resources. By exploring the effectiveness of pre-training techniques focusing on monolingual data in this context, we hope to contribute to the development of more effective MT systems for under-resourced languages, and improve access to information and services for speakers of these languages.

1.2 Motivation of the Study

Machine Translation is an essential technology for enabling communication and interaction between people who speak different languages. However, the effectiveness of machine translation systems can vary depending on the language pairs involved, and the amount and quality of available training data. While a significant amount of research has been conducted on improving machine translation for well-resourced languages, under-resourced languages like Afaan Oromo present a unique set of challenges that require special attention.

Afaan Oromo is spoken by over 40 million people primarily in Ethiopia, and it is considered one of the most under-resourced languages in the world [23]. Due to the lack of language resources, Afaan Oromo is often overlooked in the development of machine translation systems, leaving its speakers without adequate access to information and services available in other languages.

Enhancing the quality of machine translation for under-resourced languages like Afaan Oromo is crucial for promoting linguistic diversity and enabling access to information and services for speakers of these languages [23]. However, this task is challenging due to the lack of large parallel corpora and other language resources that are essential for training supervised machine translation systems.

In recent years, unsupervised learning techniques, particularly pre-training techniques like BERT and GPT, have shown great promise in learning generic language representations from large amounts of unlabeled corpus, without the need for parallel corpora [16, 24]. However, the

potential of these techniques in improving machine translation for under-resourced languages like Afaan Oromo has not been extensively studied.

Therefore, this thesis aimed to investigate the potential of unsupervised pre-training techniques in improving the effectiveness of English-Afaan Oromo machine translation, particularly in the absence of large parallel data and basic language resources. By exploring the potential of unsupervised pre-training techniques, this research aims to contribute to the body of knowledge on unsupervised learning and its application to machine translation, potentially leading to new methods for improving machine translation in under-resourced languages like Afaan Oromo.

1.3 Statement of the Problem

Machine translation has become an important area of research in natural language processing, enabling effective communication across different languages. However, machine translation has traditionally relied on parallel corpora for training, where the same text is available in both the source and target languages. The availability of parallel data is often limited, particularly for under-resourced languages like Afaan Oromo, where the lack of parallel corpora can be a major obstacle to achieving good performance in MT. Even with parallel data, supervised techniques for machine translation require a large amount of training data to achieve good performance [23, 25].

Moreover, Afaan Oromo is an under-resourced language that has very limited resources for machine translation, including a lack of even basic word embeddings. Word embeddings are fundamental tools in NLP, used to represent words as numerical vectors in order to perform various tasks such as semantic similarity and machine translation. The lack of word embeddings for Afaan Oromo further complicates the challenge of building effective MT systems for this language.

To solve the problem faced by under-resourced languages, researchers have tried to explore several methods. Considering PTLMs in previous studies, in general, three strategies have been explored. Using PTLMs to initialize the parameters of down stream models and then fine-tuning the model on downstream task, using PTLMs as context-aware embeddings for downstream task, and the use of PTLMs as teacher forcing method.

As proposed by Devlin et al. [16], the first strategy to incorporate pre-training techniques is to initialize the encoder model for downstream tasks. However, Zhu et al. [26] reported no improvements in the results after making experiments using this method on a MT task. They also experimented on XLM [25] model and concluded that there was no significant improvement when large parallel corpora is available.

Again, as context-aware embedding Zhu et al. [26] experimented the use of approach proposed by Peters et al. [27] and followed their strategy to improve the translation quality in more promising way. Focusing on this method Zhu et al. [26] proposed new approach called BERT-Fused approach and the approach is more promising in MT when compared to other discussed papers. But, this method focuses on context understanding of the source language and fails on capturing text generation task (generating accurate translation for target language). So, the above mentioned papers can't answer the question of "What happens to under resourced languages if we need the models to generate better quality text in limited parallel dataset?". No answer. Because, there is no enough parallel data for Afan Oromo language to learn generating high quality text. This creates the imbalance of understanding the source language and at the same time generating better target language in the same circumstance.

The third strategy is the use of teacher-forcing mechanism in Knowledge Distillation(KD) approach. The two previously discussed strategies both focus on language understanding. The study by Chen et al. [28] demonstrates the use of PTLMs for text generation and distilled it for downstream task (focused on text generation for target language). This approach performs better in text generation, but for rich resourced languages. This technique ignores context understanding for source language and solely focus on target generation task. This approach, in reverse to the above methods, suffers in context understanding for under resourced languages (Afaan Oromo in this case).

Based on the mentioned literature, we divided the concepts into two distinct categories: Unsupervised Language Understanding (LU) and Language Generation (LG). Existing works tend to concentrate solely on either LU or LG (by using unsupervised pre-training techniques in improving translation quality). Unsupervised pre-training techniques, such as BERT and GPT, have shown great promise in other NLP tasks, particularly in improving the quality of representations learned by neural models [16, 29]. These techniques can be used to learn generic language representations from large unlabeled data, and then fine-tune these models on specific tasks like machine translation. However, in the SOTA research, to the level of our knowledge, no studies have addressed the specific issue of improving translation quality by simultaneously considering both of these concepts. Further information are under section 2.2.

The problem this paper aimed to address is the challenge of improving the effectiveness of English-Afaan Oromo machine translation, particularly in the absence of enough parallel data and basic language resources. To tackle this problem, our research finding investigated the potential of unsupervised pre-training techniques in improving the quality of Afaan Oromo machine translation. By exploring this approach, this paper contributes to the body of knowledge on incorporation of unsupervised learning techniques and their application to neural machine translation (NMT), potentially leading to new methods for improving NMT in under-resourced

languages like Afaan Oromo.

1.4 Research Questions

The following questions are the top questions that this paper is going to address.

1. How can we marry the concepts of unsupervised language understanding and language generation to get better translation quality?
2. How does the incorporation of pre-training techniques, such as BERT and GPT, in transformer model improve the quality of English-Afaan Oromo machine translation?

1.5 Objective of the Research

1.5.1 General Objective

The general objective of this study is to explore the potential of marrying unsupervised Language Understanding (LU) and Language Generation (LG) techniques and incorporating them to Transformer Model to improve the quality of English-Afaan Oromo machine translation using monolingual data.

1.5.2 Specific Objectives

The specific objective of this research is:

- To explore the use of marrying both unsupervised language understanding and language generation concepts in improving translation quality.
- To investigate the effectiveness of incorporating unsupervised pre-training techniques to transformer model for improving the quality of English-Afaan Oromo machine translation.
- To identify the performance of pre-training techniques with transformer model to NMT with monolingual data.
- To develop a model which incorporates unsupervised LU and LG for better translation quality.

1.6 Contribution of the Study

This Thesis makes several contributions to the field of English-Afaan Oromo machine translation and under-resourced language processing:

Dataset Collection: The study involved the collection of two datasets, including an English-Afaan Oromo parallel dataset and an Afaan Oromo monolingual dataset. These datasets contribute to the availability of resources for future research and development in machine translation for Afaan Oromo.

Methodology Exploration: The study explored the marriage of two approaches, unsupervised language understanding using BERT and language generation using GPT, to enhance the Transformer model's performance. The marriage of these unsupervised language understanding and language generation concepts within the transformer model presents a unique and innovative approach to machine translation. By combining the fine-tuned BERT model for language understanding with the fine-tuned GPT-2 model for language generation, the proposed system demonstrated the ability to produce high-quality translations in under-resourced languages in limited parallel dataset.

1.7 Scope and Delimitation of the Study

The scope of this research is focused on investigating the potential of unsupervised pre-training techniques in improving machine translation for English-Afaan Oromo. The research involves experimenting with different pre-training techniques on large amounts of unlabeled data, and evaluating their effectiveness in improving the performance of machine translation systems. The research also involves evaluating the performance of machine translation systems on different text genres and domains, to understand the generalizability of the proposed approach.

The delimitation of this research includes the following:

- The research focuses exclusively on English-Afaan Oromo machine translation, and does not explore other language pairs.
- The research assumes the absence of large parallel data and basic language resources for Afaan Oromo, and does not explore machine translation approaches that rely on these resources.
- The research does not explore the potential of other unsupervised learning techniques beyond monolingual pre-training techniques, such as clustering and auto-encoders.
- The evaluation of the machine translation systems is limited to standard automatic evaluation metrics (BLEU score, ChrF score, and TER), and does not involve human evaluation.

By setting these delimitations, the research is focused on investigating the potential of unsupervised pre-training techniques in improving machine translation for under-resourced languages, specifically on Afaan Oromo, and provide insights into the effectiveness of these techniques in addressing the unique challenges faced by machine translation systems for under-resourced

languages.

1.8 Structure of the document

The rest of this thesis is structured as follows. Following the Introduction section, a section on Literature Review, which critically reviews the existing literature on machine translation, with a particular focus on English-Afaan Oromo machine translation and pre-training in transformer models for unsupervised learning will be discussed. It aims to identify gaps in the literature that this paper will address. Following the Methodology section, which outlines the research design, model description, data collection methods, and evaluation techniques that will be used to address the research questions and objectives. It includes a detailed discussion of the pre-training approach for unsupervised learning and the specific transformer model architecture that will be used in the study and evaluation metrics used to evaluate the effectiveness of the proposed methodology. Under Result and Analysis, the experimental result of the findings will be discussed. Finally, we discuss the conclusion and recommendation for future work and further study.

Chapter Two

2. Literature Review

2.1 Introduction

In this section, we critically discuss the overview of Afaan Oromo language and an overview of Machine Translation (MT) including its historical background, approaches, and evaluation metrics. Furthermore, we will discuss different available related works with our conclusion idea.

2.1.1 Overview of Afaan Oromoo Language

Afaan Oromo, also known as Oromiffa, is an Afro-Asiatic language spoken by the Oromo people in Ethiopia, Kenya, and Somalia. It is the sixth most widely spoken language in Africa, with over 40 million speakers worldwide [30]. The language has a rich oral tradition and is spoken in various dialects. Afaan Oromo is a Cushitic language that originated in the Horn of Africa and is now spoken primarily in Egypt, Ethiopia, Kenya, Somalia, and the United Republic of Tanzania. The majority ethnic group in Ethiopia, the Oromo, makes up more than 40% of the country's total population [31].

2.1.1.1 Afaan Oromo Writing System

In the 1970s, the Oromo Liberation Front (OLF) developed a modified Latin-based script that is now widely used. This script uses 33 letters to represent the sounds of the language, including five vowels and 28 consonants. Five vowels and eighteen basic consonants (dubbifamaa) are used to write the borrowed or foreign words. Three additional consonants are also used. Letters with two consonants combined into them are known as double consonant letters. The English alphabet's capital and little letters, known as Qubee, are its distinguishing features. Vowels are sound creators and sound by themselves, just like in English. Afaan Oromo vowels can be classified as short or long. Because 'p', 'v', and 'z' are not found in any native Afaan Oromo words, the basic consonant also does not contain them. However, these characters are utilized when referring to foreign nouns like "Video" (Viidiyoo in Afaan Oromo).

An apostrophe (') is a punctuation mark that can also be used to indicate a sound in addition to these other symbols. It refers to the hudhaa sound, which is heard in the words re'ee and fal'aana [32]. Table 1 shows the Afaan Oromo writing system, as taken from [33].

Table 1. *Afaan Oromo Writing Scripts*

A	B	C	CH	D	DH
E	F	G	H	I	J
K	L	M	N	NY	O
P	PH	Q	R	S	SH
T	TS	U	V	W	X
Y	Z	ZH			

The vowels in the Afaan Oromo language are denoted by the five letters a, e, i, o, and u. In Oromia, vowels are often pronounced in the same way. When emphasized, these vowels can be opened (deemi, "go", or nyaadhu "eat") or closed (bada, rafi, etc.). Every word in Afaan Oromo is pronounced powerfully because the vowels are always pronounced sharply and clearly. For example:

A Arba, Farda, Haadha
E Gannaalee, Waabee, Noolee, Roobalee, collee
I Arsii, Laali, Rafi, Lakki, Sirbi
O Oromoo, Cilaaloo, Haroo, caancco, Danbidoolloo
U Ulfaadhu, Guddadhu, dubbadhu, Ituu

2.1.1.2 Afaan Oromo Morphology

Afaan Oromo has a complex system of morphology that is essential to understanding the language like many other languages. Morphology is the study of the structure of words and how they are formed. Afaan Oromo has a rich system of morphology that involves the use of prefixes, suffixes, infixes, and stem modifications to create new words and convey different meanings [34].

One important feature of Afaan Oromo morphology is the use of gender and number markers [35]. These markers are attached to nouns and indicate whether the noun is masculine or feminine, and whether it is singular or plural. For example, the noun "Kitaaba" means "book" in Afaan Oromo. If we want to indicate that there are multiple books, we can add the plural marker "-ota" to get "Kotaabota". If we want to indicate that the books belong to a woman, we can add the feminine marker "-shee" to get "Kitaabotashee". Table 2 shows the most common Afaan Oromo prular forms.

Table 2. *Afaan Oromo most common prular morphs*

Base Forms	**Inflected Forms**	**English meaning**
Karra	Karr-oota	Gates
Laga	Lag(g)-een	Rivers
Jabbii	Jabbii-lee	Calfs
Jaarsa	Jaars-olii	elders
Barmaata	Barmaat-ilee	Habits
Qaama	Qaam-olee	Bodies

Another important aspect of Afaan Oromo morphology is the use of verbal inflection to indicate tense, aspect, and mood. Verbs in Afaan Oromo have a variety of inflectional suffixes that are added to the stem to indicate different meanings. For example, the suffix "-ti" is used to indicate the past tense, while the suffix "-n" is used to indicate the present tense. The suffix "-uu" is used to indicate the imperative mood, while the suffix "-f" is used to indicate the subjunctive mood.

Afaan Oromo also has a rich system of derivational morphology, which involves the use of affixes to create new words from existing ones. For example, the prefix "hin-" and "-iin" can be added to a verb to create a negative form, as in "hin-dubbis-iin" (not read).

In addition to these affixes, Afaan Oromo also has a system of infixes, which are inserted into the middle of a word to change its meaning. For example, the infix "-kko-" can be added to a verb stem to indicate that the action is repeated, as in "kokkolfe" (he loughs repeatedly).

Afaan Oromo morphology is a complex and rich system that plays an important role in the language. By understanding the various affixes and inflections, speakers of Afaan Oromo are able to create new words and convey a wide range of meanings.

2.1.1.3 Afaan Oromo Sentences

The language has its own unique sentence structure and grammar rules. In Afaan Oromo, a sentence usually follows a subject-object-verb (SOV) order, where the subject comes first, followed by the object and then the verb. For example, the sentence "Chala hit Tola." in Afaan Oromo would be "Caalaan Tolaa rukute." where "Caalaa" is the subject, and "Tolaa" is the object, and "rukute" is the verb (meaning "hit").

sentences in Afaan Oromo can be classified based on their structure as simple, complex, compound, and compound-complex [36].

Simple Sentence: A simple sentence is a sentence that contains only one independent clause. It expresses a complete thought and can stand alone as a sentence. For example, "Abbaan koo taa'aa jira." meaning "My father is sitting."

Complex Sentence: A complex sentence is a sentence that contains one independent clause and one or more dependent clauses. The dependent clause cannot stand alone as a sentence and is usually introduced by a subordinating conjunction such as "yoom" meaning "when," "yoo" meaning "if," or "sababni isaa" meaning "because." For example, "Osoon kaleessa gaafa ati dhufte achi jiraadhee, silaa nagaa sin gaafadha ture." meaning "If I were there yesterday when you came, I would have greeted you."

Compound Sentence: A compound sentence is a sentence that contains two or more independent clauses joined by a coordinating conjunction such as "fi" meaning "and," "garuu" meaning "but," or "yookiin" meaning "or". For example, "Oromoon gama biraatiin aadaa fi duudhaa badhaadhaa, kan ittiin boonu fi kunuunsuuf carraaqu qaba." meaning "Oromos, on the other hand, have a rich culture and tradition, which they are proud of and which they strive to preserve."

Compound-Complex Sentence: A compound-complex sentence is a sentence that contains two or more independent clauses and one or more dependent clauses. It combines the features of both compound and complex sentences. For example, "Kaleessa gaafa dhufte osoon jiraadhee, silaa nagaa sin gaafadha ture, garuu abbaan koo yeroo sanatti bahee ture." meaning "When you came yesterday, if I were there, I would have greeted you, but my father had gone out at that time."

2.1.2 Overview of Machine Translation

Natural Language Processing (NLP) has been an active research field for several decades, and it has been used in various applications such as MT, information retrieval, sentiment analysis, and speech recognition. One of the major challenges in NLP is MT, which involves translating text from one language to another. Machine translation is the process of automatically translating text from one language to another. It has been an active area of research since the 1950s [37].

Translation is more than just word-for-word translation. It takes extensive knowledge of grammar, syntax (sentence structure/word order), semantics, etc., in the source and target languages as well as familiarity with each local region in which syntax and semantic mean of sentence structure and meanings respectively. A translator must interpret and analyze all elements of a text and know how each word may influence another [38].

Machine translation offers several advantages, making it an efficient tool for individuals and

businesses who need to communicate across languages. Using MT, large volumes of text or speech can be quickly translated, enabling organizations to communicate with their global audience in real-time. This can be particularly useful in industries such as e-commerce, where businesses need to provide accurate and timely translations of product descriptions and customer reviews to international customers [39]. Another advantage of MT is cost-effectiveness. Hiring human translators can be expensive, especially for businesses that require large volumes of content to be translated. Machine translation, on the other hand, offers a more affordable solution, making it possible to translate vast amounts of content at a fraction of the cost of hiring a human translator. This can be particularly beneficial for small and medium-sized businesses that operate on tight budgets [39].

Machine translation also offers a high degree of accessibility, breaking down language barriers and making content more accessible to people who speak different languages. This can be particularly beneficial for organizations that operate in multilingual environments or that provide services to global audiences. By making content accessible in multiple languages, businesses can reach a wider audience and create a more inclusive environment. The study by Brynjolfsson et al. [40] estimate that the introduction of MT on eBay has led to a 17.5% increase in international trade, equivalent to an additional $2 billion in trade volume.

Machine translation models can be customized to suit specific industries or domains. This means that translations can be tailored to the needs of a particular business or organization, making it possible to generate translations that are accurate and relevant to a particular industry or domain. This can be particularly beneficial for businesses that operate in specialized industries, such as healthcare or legal services, where accurate translations of technical terminology are essential.

2.1.3 History of Machine Translation

The history of MT dates back to the 1940s, when researchers began exploring the possibility of using computers to automatically translate languages. In 1949, Warren Weaver [41], a mathematician and scientist at the Rockefeller Foundation, published an influential article titled "Translation" in which he outlined a vision for machine translation. Weaver argued that translation was a problem of "interlingua" or "hidden" language, and that a computer could be used to mediate between the source language and the target language by first translating the source language into this interlingua, and then translating the interlingua into the target language.

In the 1950s, researchers began developing the first MT systems, including the Georgetown-IBM experiment and the Systran system [42]. These early systems used rule-based approaches, in which linguistic rules were manually coded into the system to generate translations. However,

these systems were limited by the complexity of natural language and the difficulty of capturing all of its nuances and irregularities.

In the 1960s and 1970s, researchers began exploring statistical approaches to machine translation, in which the system learned to translate by analyzing large corpora of parallel texts. One notable example of this approach was the ALPAC report, which evaluated the state of MT research and found that statistical approaches showed more promise than rule-based approaches [43, 44].

In the 1980s and 1990s, researchers continued to refine statistical approaches to MT, and also began exploring the use of neural networks for MT [8]. However, progress was slow due to the computational resources required for training and testing these systems.

In the 2000s and 2010s, the availability of large-scale parallel corpora and advances in computing power led to significant improvements in MT performance. The introduction of phrase-based and then NMT models, which use neural networks to learn to translate at the sentence or sub-sentence level, has led to state-of-the-art performance on a variety of language pairs [45]. The current state of the art in MT is dominated by NMT models. Various techniques have been developed to improve NMT performance, including attention mechanisms, which allow the decoder to focus on different parts of the source sentence during decoding, and subword modeling, which allows the model to handle rare or out-of-vocabulary words [46].

Overall, the current state of the art in MT is characterized by the dominance of NMT models, which have shown remarkable progress in recent years and are the focus of ongoing research and development efforts. However, despite its impressive performance, NMT still faces challenges such as the difficulty of translating low-resource languages or translating between languages with vastly different linguistic structures.

2.1.4 Approaches of Machine Translation

There are three main approaches of machine translation: rule-based, statistical, and neural machine translation. Rule-based machine translation uses linguistic rules and dictionaries to translate text, while statistical machine translation uses statistical models to learn the translation patterns from a large amount of parallel corpora. Neural machine translation, on the other hand, uses neural networks to model the translation process, and has become the dominant approach in recent years.

2.1.4.1 Rule-based Machine Translation

Rule-based Machine Translation (RBMT) is an approach to machine translation that relies on explicitly written rules to translate text from one language to another. It involves creating a set of linguistic rules and grammar patterns that specify how words and phrases should be translated from the source language to the target language. These rules are typically created by linguists and language experts, and they take into account the grammar, syntax, and vocabulary of both languages [44].

The RBMT approach involves breaking down sentences into their component parts and then using the rules to translate each part. For example, a rule might specify that a verb in the source language should be translated into a particular tense in the target language. RBMT systems can also use dictionaries and databases to store information about words and phrases in the source and target languages, which can help improve the accuracy of the translations [44].

One of the major challenges of RBMT is the complexity of the rules required to accurately translate natural language. Language is constantly evolving and changing, and creating rules that can account for all the nuances and idiosyncrasies of a language can be difficult. Additionally, RBMT systems may struggle with ambiguity, such as when a word has multiple possible meanings, or when a sentence can be interpreted in multiple ways. Another challenge of RBMT is that it typically requires a large amount of human input and expertise to create and maintain the rules. This can be time-consuming and expensive, and it may not be practical for all languages or for languages with limited resources.

2.1.4.2 Statistical Machine Translation

Statistical Machine Translation (SMT) is a machine translation approach that uses statistical models to automatically learn how to translate text from one language to another [8]. Unlike rule-based machine translation, SMT does not rely on pre-written rules, but instead uses statistical models that are trained on large parallel corpora (i.e., a set of texts in the source language and their translations in the target language).

The basic idea behind SMT is to use these parallel corpora to learn patterns and statistical relationships between words and phrases in the source and target languages [8]. These patterns can then be used to generate translations for new sentences or texts. SMT systems typically use two main components: a translation model, which determines the most likely translation for a given sentence or phrase, and a language model, which helps to generate fluently worded translations by predicting the likelihood of certain sequences of words in the target language.

One of the main challenges of SMT is the need for large amounts of parallel corpora in order to

train accurate translation models [47]. For under-resourced languages, which may have limited amounts of parallel data available, this can be a significant obstacle. Additionally, even when parallel corpora is available, it may not be of high quality, which can lead to inaccuracies in the resulting translations. Another challenge of SMT is that it can struggle with translating idiomatic expressions or other forms of language that are not easily captured by statistical patterns. For example, SMT may have difficulty with idiomatic expressions like "kick the bucket" or "break a leg" that have non-literal meanings. while Statistical Machine Translation has been successful for a number of language pairs, it does come with a number of challenges, particularly for under-resourced languages. In recent years, the development of NMT has shown promise for addressing some of these challenges by leveraging deep learning techniques to improve translation accuracy and reduce the need for large amounts of parallel data.

2.1.4.3 Neural Machine Translation

Neural Machine Translation (NMT) is a machine translation approach that uses artificial neural networks (ANN) to learn how to translate text from one language to another [45]. NMT represents a significant departure from earlier machine translation approaches such as rule-based and statistical machine translation. Instead of relying on pre-defined rules or statistical models, NMT systems are trained to learn patterns in the data, allowing them to generate more natural and fluent translations.

The basic idea behind NMT is to use an encoder-decoder architecture that takes a source sentence in one language and generates a target sentence in another language [48]. The encoder component processes the source sentence and converts it into a fixed-length vector representation, which is then used by the decoder component to generate the corresponding target sentence.

NMT has several advantages over earlier machine translation approaches. For one, NMT can handle longer and more complex sentences than earlier approaches. Additionally, NMT is able to learn from the data and generate more natural-sounding translations, with better fluency and accuracy [49]. Neural machine translation has shown significant improvements in translation quality over the traditional methods. In the encoder-decoder architecture, the input sentence is first encoded into a fixed-length vector, then decoded into the output sentence [13]. Attention mechanisms have also been introduced in NMT to allow the model to focus on different parts of the input sentence during the decoding process [46, 50].

One of the main challenges of NMT is the need for large amounts of parallel corpora to train the neural network [51]. While this is less of an issue for widely spoken languages with large amounts of data available, it can be a significant challenge for under-resourced languages. Additionally, NMT models can be computationally expensive and require significant computing

resources to train and run. Another challenge of NMT is the issue of domain adaptation. While NMT models can learn from general-purpose data, they may struggle when translating text in specialized domains such as medical or legal documents. Domain adaptation techniques, such as fine-tuning or transfer learning, can help to address this challenge.

2.1.5 The Transformer Model

Before Transformer model was introduced, previously sequence-to-sequence models were such as recurrent neural networks (RNNs) [52] with long short-term memory (LSTM) [53] units or gated recurrent units (GRUs)[54]. While RNNs had been widely used for tasks like machine translation, they suffered from computational inefficiency due to sequential computation, making them slow for long sequences. Additionally, RNNs struggled to capture long-term dependencies in sequences, as information from earlier time steps degraded as it propagated through time. The lack of a mechanism to effectively capture global dependencies across the entire sequence was also a limitation of RNNs [46].

The Transformer model addressed these limitations by introducing the concept of self-attention [46]. Self-attention allows the model to focus on different parts of the input sequence simultaneously, enabling highly parallel computation. This parallelism significantly improved the efficiency of training and inference compared to sequential models like RNNs. Moreover, the self-attention mechanism enabled the Transformer model to model dependencies between any pair of positions in the input sequence, regardless of their temporal distance. This capability allowed the model to capture long-term dependencies more effectively. By attending to all positions in the input sequence, the Transformer model could also capture global dependencies and consider the context of each token in relation to the entire sequence. This global context improved the model's ability to generate accurate translations by incorporating information from the entire input sequence. The Transformer model consists of the following core components:

2.1.5.1 Input Embeddings

The input to the Transformer model is a sequence of tokens, denoted as $X = x_1, x_2, ..., x_n$, where n is the sequence length. Each token is first transformed into a dense vector representation called an embedding. This is typically achieved using an embedding matrix E. The embedding of token x_i is given by $e_i = E[x_i]$.

2.1.5.2 Positional Encodings

To incorporate positional information, the Transformer model adds positional encodings to the input embeddings. The Transformer model employs sinusoidal functions at various frequencies to create position encodings across the sequence. As a result, neighboring elements in the

sequence are assigned similar position encodings due to the nature of these sine and cosine functions. A positional encoding matrix PE is created, where each row corresponds to a position in the sequence. The positional encoding for position pos and dimension d is calculated as follows:

$$PE_{(pos,2d)} = sin(pos/10000^{2d/d_{model}})$$
$$PE_{(pos,2d+1)} = cos(pos/10000^{2d/d_{model}})$$

Here, d_{model} represents the dimensionality of the model. The positional encodings are then added element-wise to the input embeddings.

2.1.5.3 Encoder

The encoder in the Transformer model consists of multiple layers, each composed of two sub-layers: the multi-head self-attention mechanism and the position-wise feed-forward neural network.

2.1.5.3.1 Multi-head Self-Attention The self-attention mechanism allows the model to weigh the importance of different tokens when encoding information. The Transformer employs scaled dot-product attention [46]. Given an input sequence X and its corresponding embeddings E, the self-attention operation can be defined as follows:

$$Attention(q, K, V) = softmax(\frac{qK^T}{\sqrt{d_k}})V \tag{2.1}$$

In this equation, q, K, and V are obtained by linearly transforming the input embeddings E:

$$q = W_q E$$
$$K = W_K E$$
$$V = W_V E$$

Here, W_q W_K, and W_V are learnable weight matrices specific to the self-attention mechanism. The output of the self-attention mechanism is the weighted sum of the values V, which captures the contextual information for each token.

2.1.5.3.2 Position-wise Feed-forward Neural Network After the self-attention mechanism, a position-wise feed-forward neural network is applied independently to each token in the sequence. It consists of two linear transformations:

$$FFN(x) = max(W_1 x + b_1, 0)W_2 + b_2 \tag{2.2}$$

Here, W_1, b_1, W_2, and b_2 are learnable weight matrices and biases. This operation introduces non-linearity and enables the model to capture complex relationships between tokens.

2.1.5.4 Decoder

The decoder in the Transformer model also consists of multiple layers, each with two sub-layers: masked multi-head self-attention and encoder-decoder attention.

2.1.5.4.1 Masked Multi-head Self-Attention In the decoder, the self-attention mechanism employs masking to prevent positions from attending to subsequent positions [46]. This ensures that during the generation of each token, only previous tokens are considered. The masked self-attention operation is similar to the one used in the encoder, but with a masking step applied to the attention weights.

2.1.5.4.2 Encoder-Decoder Attention The encoder-decoder attention mechanism allows the decoder to focus on relevant parts of the input sequence from the encoder. It attends to the encoder's outputs and combines the information with the decoder's own representations. The operation can be defined as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_{k_{enc}}}})V_{enc} \tag{2.3}$$

Here, Q and V are obtained by linearly transforming the decoder's output embeddings, while K and V_{enc} are obtained by linearly transforming the encoder's output embeddings.

2.1.5.5 Output Layer

The output layer of the Transformer model is a softmax layer that predicts the probability distribution over the target vocabulary. The model learns to generate the most likely translation by maximizing the likelihood of the correct target token at each time step.

2.1.6 Unsupervised Learning

Unsupervised learning is a category of machine learning where the model learns patterns and structures in data without explicit labels or guidance. Unlike supervised learning, which relies on labeled examples to make predictions, unsupervised learning aims to discover hidden patterns, relationships, and structures within the data itself. This type of learning is particularly useful when dealing with large and unannotated datasets, as it allows the model to extract meaningful insights without human intervention.

Pre-training is one of unsupervised learning method which boosted the performance of different machine learning tasks including machine translation. Pre-training is a method employed within the broader framework of unsupervised learning. In unsupervised learning, the goal is to extract meaningful patterns and structures from unlabeled data. Pretraining, specifically, is a technique that leverages large amounts of unannotated data to learn useful representations or features before applying supervised learning or fine-tuning on a specific task.

The motivation behind pre-training arises from the insight that learning generic features from vast amounts of unlabeled data can provide a good initialization point for subsequent supervised learning tasks. By first training a model on unsupervised tasks, it can learn to capture underlying patterns and structures in the data, which can then be transferred and fine-tuned for downstream tasks with limited labeled data.

There have been various pre-training approaches developed to learn effective representations across different domains. Some notable pretraining methods in machine translation specifically considering low-resource languages include: Cross-lingual Transfer Learning, Multilingual Pre-training, and Monolingual Pre-training approaches.

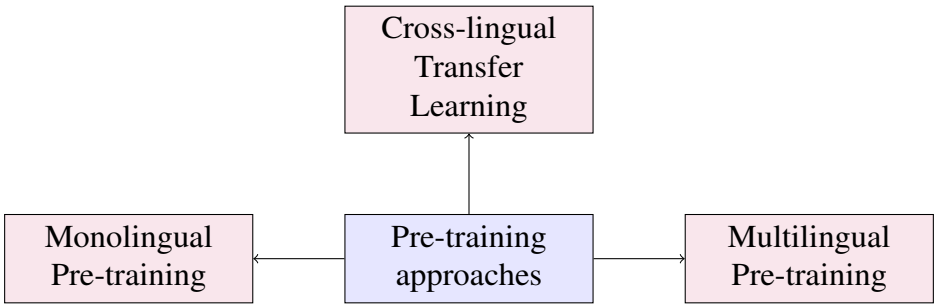


Figure 1. Pre-training approaches for under-resourced machine translation

2.1.6.1 Cross-lingual Transfer Learning

Cross-lingual transfer learning involves pre-training an NMT model on a high-resource language pair with abundant parallel data. The model learns general language representations and translation knowledge during pre-training, which can be transferred to improve translation quality for low-resource languages. This technique enables the model to leverage shared linguistic patterns and translation cues across languages, enhancing its performance on low-resource language pairs. Notable works, such as Artetxe [55], have demonstrated the effectiveness of unsupervised pre-training followed by supervised fine-tuning using cross-lingual transfer learning. Kim [56] proposes techniques to transfer a pretrained NMT model to a new language without shared vocabularies, achieving improved translation performance. Ji [57] focuses on zero-shot translation and introduces cross-lingual pre-training methods to enable smooth transitions between different languages, outperforming pivot-based and multilingual NMT approaches. Kim [58]

explores pivot-based transfer learning using parallel corpora involving a pivot language, demonstrating significant improvements in source-target translation. Ren [59] presents a novel explicit cross-lingual pre-training method for unsupervised machine translation, incorporating explicit cross-lingual training signals and achieving improved translation performance. Imankulova [60] suggests a multilingual multistage fine-tuning approach that leverages out-of-domain data to improve translation quality in low-resource scenarios. Kocmi [61] presents a simple transfer learning method where a "parent" model trained on a high-resource language pair is continued on a low-resource pair, resulting in significant improvements. Overall, these papers highlight the effectiveness of different pre-training methods for cross-lingual transfer learning in NMT, showcasing improved translation performance and addressing challenges in low-resource and zero-shot translation scenarios.

2.1.6.2 Multilingual Pre-training

Multilingual pre-training involves training an NMT model on large-scale multilingual corpora, encompassing diverse languages. The model learns from the shared linguistic properties and translation patterns across multiple languages during pre-training. This technique enables the model to capture universal language representations and transferable translation knowledge. Notable examples include the XLM-R model proposed by Conneau [62], which demonstrated impressive performance on low-resource NMT tasks through unsupervised cross-lingual representation learning. Lakew [63] demonstrates the effectiveness of multilingual NMT in translating low-resourced languages by creating shared semantic representations across multiple languages. Dabre [64] explores multistage fine-tuning configurations and highlights the utility of multi-parallel corpora for transfer learning in low-resource NMT. Zhang [65] proposes a new pre-training method that incorporates bilingual data to improve the generalization ability of pre-training models and enhance translation performance. Liu [66] presents a continual pre-training framework to adapt multilingual pre-trained models to unseen languages, achieving consistent improvements in fine-tuning performance for low-resource translation pairs.

2.1.6.3 Monolingual Pre-training

Monolingual pre-training focuses on utilizing large amounts of monolingual data available for the target low-resource language. The NMT model is pre-trained on this monolingual data to learn general language representations, contextual understanding, and language generation capabilities. This pre-training step aims to improve the model's language understanding abilities, including vocabulary, grammar, and context. Raffel [67] introduced a text-to-text transformer model that achieved promising results in unsupervised machine translation through monolingual pre-training. Zhang [65] proposes a pre-training method that incorporates bilingual data, showing improved generalization ability and performance in translation models. Liu [66] presents a continual pre-training framework that adapts a multilingual pre-trained model to unseen

languages, consistently improving fine-tuning performance. Gangi [68] suggests using word embeddings trained on monolingual data to enhance NMT models, demonstrating positive results in low-resource scenarios. Tonja [69] explores the use of source-side monolingual data to improve NMT systems for low-resource languages, achieving significant BLEU score improvements through self-learning and fine-tuning approaches. These papers collectively highlight the effectiveness of various monolingual pre-training methods in enhancing low-resource neural machine translation.

2.1.7 Evaluation Metrics

There are several evaluation metrics used to measure the performance of MT systems. Here are some of the most commonly used evaluation metrics for Machine Translation:

BLEU score: The Bilingual Evaluation Understudy (BLEU) score is a popular metric for evaluating machine translation systems. BLEU measures how well the generated translations match human translations, based on n-gram precision. It computes a score between 0 and 1, with higher scores indicating better translation quality [70]. BLEU score is based on the n-gram precision, which measures the percentage of n-grams (contiguous sequences of words) in the machine-generated translation that also appear in the reference translations. BLEU score is easy to compute and widely used, but it has some limitations, such as not taking into account the meaning of the words.

METEOR: The Metric for Evaluation of Translation with Explicit ORdering (METEOR) is another widely used evaluation metric for machine translation. METEOR uses a combination of n-gram precision, recall, and a measure of how well the generated translation matches the reference translations. METEOR also takes into account the synonyms, paraphrases, and word order differences between the generated translation and the reference translations. The score ranges from 0 to 1, with a score of 1 indicating a perfect match between the machine-generated translation and the human reference translation [71].

TER: Translation Error Rate (TER) measures the number of edits required to transform the generated translation into the reference translation. TER scores range from 0 to 1, with lower scores indicating better translation quality [72].

NIST: The National Institute of Standards and Technology (NIST) metric is similar to BLEU, but it uses a different weighting scheme to measure the n-gram precision of the generated translations. NIST scores range from 0 to 1, with higher scores indicating better translation quality [73]. NIST is widely used in machine translation evaluations, especially for government-sponsored evaluations.

HTER: Human Target Error Rate (HTER) is an evaluation metric that measures the distance between the generated translation and the target (i.e., reference) translation in terms of the number of edits required to make the generated translation indistinguishable from a human translation. HTER scores range from 0 to 1, with lower scores indicating better translation quality [74].

It's worth noting that these evaluation metrics have their own strengths and weaknesses, and they may not always provide a complete picture of the quality of the translations produced by an MT system. Therefore, it's recommended to use multiple evaluation metrics and to also conduct manual evaluations to assess the overall translation quality.

2.2 Related Works

Under this section we will see other related works in two folds. Firstly we will discuss different works explored on Ethiopian languages including their strength and weaknesses. And secondly we will investigate different approaches applied on MT using pre-training methods considering under-resourced languages.

2.2.0.1 Machine Translation Systems for Ethiopian languages

Different works have been conducted using different approaches to different Ethiopian languages to apply it on Machine Translation. These approaches include Rule based MT, Statistical MT and Neural MT. From these approaches NMT is the dominant one. After Vaswani et al. [46] introduced the transformer model, which has become the dominant architecture for NMT due to its ability to handle long-range dependencies and its superior performance compared to earlier models, some works are conducted using this approach on Ethiopian Languages.

Ashengo, Aga, and Abebe [75] investigated a new approach for translating English text to Amharic text using a combination of context-based machine translation (CBMT) and a recurrent neural network machine translation (RNNMT). The paper shows that the combinational approach on the English-Amharic language pair yields a performance improvement over the simple NMT, while no improvement is seen over CBMT for a small dataset. The impact of the dictionary used by CBMT on the overall performance of the approach is also assessed. Arfaso [76] used a CNN-based bi-directional text-based MT for language pairings of English and Afaan Oromo. By using CNN on translations between these language pairs, the author's first goal was to improve the earlier work on English to Afaan Oromo MT by making the translation bi-directional. Dereje [77] used hybrid artificial neural machine translation techniques to carry out his research work from English to Afaan Oromo. He suggested a concept that is put into practice by dividing it into three layers: an encoder layer, a hybrid layer, and a decoder layer. The proposed model

improves the translation quality by addressing issues such as lack of decoding mechanism, lack of translation for long sentences, and out-of-vocabulary words leading to unknown word output. But, the mentioned papers suggested that their method were not sufficient to handle long range dependencies between terms in a sentence.

Meron [78] worked on developing an attention-based Amharic to Afaan Oromo NMT system using an Encoder-Decoder model with a Bi-directional Gated Recurrent Unit (BGRU). The non-attention based system with basic encoder-decoder architecture has limitations when it comes to longer sentences, as the inter-dependency of words increases. The paper also notes that prior to this work, there was no other NMT system translating Amharic to Afaan Oromo. The performance of the attention-based system is compared to that of the non-attention based system. The result of the proposed appproch shows that the attention mechanism can handle longer sentences when compared to non-attention approaches. But, the GRU and RNN architectures are inherently sequential, as each time step relies on the previous time step. This sequential nature makes it challenging to fully parallelize these models for improved translation. In contrast, the Transformer model has a distinct advantage with its ability to parallelize computations across its layers, allowing it to outperform GRU and RNN models in terms of translation efficiency.

Waldeyes [33] employed a bi-directional MT strategy for the Recurrent NN (Bi-encoder and decoder Gated Recurrent Units (GRU) and Bi-encoder and decoder Long Short-Term Memory (LSTM)) with attention and Transformer model for the language pairs of English and Afaan Oromo. This paper introduced the transfer approach rather than using conventional NMT with simple attention. The result of this paper is promising for English-Afaan Oromo pair but the problem is that it is difficult to get large parallel corpus to make this transformer model work for under-resourced languages such as Afan Oromo. Considering this, Tonja et al. [69] explores the use of source-side monolingual data to improve NMT systems for low-resource languages, particularly using Wolaytta–English translation using both authentic and synthetic datasets achieving significant BLEU score improvements through self-learning and fine-tuning approaches. Again here, they used back-translation technique by using synthetic dataset. This is common problem for such models for generalizability and applicability of the models on real-world application.

Based on the literature provided, it appears that the NMT research on Ethiopian languages faces significant challenges due to data scarcity. The largest available dataset consists of only 25,000 parallel sentences [79], which poses a limitation for existing methods. This emphasizes the critical need for focused attention and concerted efforts in this area. By addressing the data scarcity problem and expanding the available resources, we can make significant advancements in machine translation for Ethiopian languages. Table 3 summarizes some SOTA related works conducted on some Ethiopian languages.

Table 3. *Summary of SOTA related works on NMT for Ethiopian Languages*

No	Authors (Year)	Title	Approach	Dataset Used	Language	Result	Remark
1	Yeabsira Asefa (2020) [75]	CBMT with RNN for English to Amharic Translation.	Combination of RNNMT and CBMT	A parallel 8,603 phrases and sentences from the New Testament of the Bible were used as a corpus.	English-Amharic	The CBMT outperformed the NMT on average by 14.23 BLEU points (18.3% for English and 10.16% for Amharic).	He used small Parallel dataset, Domain specific
2	Arfaso Birhanu (2019) [76]	Bi-Directional English to Afaan Oromo MT using CNN	CNN	5,550 sentences from religious texts, conversational novels, internet, Oromia region revenue, and the health sector.	English-Afaan Oromo	24.37 BLEU score for English to Afaan Oromo, and 23.18 for Afaan Oromo to English.	Small corpus, Parallel dataset, Doesn't show effect on large sentences.
3	Dereje Seifu (2019) [77]	Hybrid ANN using deep learning techniques English to Afaan Oromo	SMT, LSTM and HNMT.	A corpus of 13,300 parallel documents from religious websites, the FDRE criminal code, and the Council of Oromia regional state.	English-Afaan Oromo	The method led to a gain of up to 2.4 BLEU points on test sets that translated English into Afaan Oromo.	Small corpus, parallel datataset, low BLUE result

Continues...

Table 3 – *Continues...*

No	Authors (Year)	Title	Approach	Dataset Used	Language	Result	Remark
4	Meron Gashaw (2021) [78]	Attention-based Amharic-to-Afaan Oromo Neural Machine Translation	NMT with Attention	A total of 11,727 sentence pairs taken from the Holly Bible and other religious and legal sources.	Amharic-Afaan Oromo	Attention-based system received a score of 67.82% over non-attention-based system which is 61.49%.	Small parallel corpus, domain specific
5	Waldeyes Adere (2022) [33]	Bi-Directional English to Afaan Oromo NMT	LSTM, GRU and Transformer	They used 21,323 parallel corpora collected from Holy Bible and previously collected parallel corpus by Yitayew [80].	English-Afaan Oromo	The BLUE score increased from 6.48 to 16.40 using the Transformer model over others.	Small parallel dataset
6	Atnafu Lambebo Tonja (2023) [69]	Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data	LSTM, Bi-LSTM and Transformer	They utilized 30,495 Wolaytta–English parallel datasets from the Tonja et al. [79] and 40k monolingual Wolaytta dataset collected from several sources.	Wolaytta – English	The self-learning approach yielded +2.7 and +2.4 BLEU score improvements for Wolaytta to English and English to Wolaytta translations.	Small parallel dataset and small monolingual data.

2.2.0.2 Pre-training approaches on NMT

In recent years, unsupervised learning and pre-training in transformer models have shown promise in improving NLP tasks for low-resource languages. Several studies have investigated the use of unsupervised learning for low-resource machine translation. Vaswani et al. [46] introduced the transformer model, which has become the dominant architecture for NMT due to its ability to handle long-range dependencies and its superior performance compared to earlier models such as RNNs and LSTMs. Devlin et al. [16] proposed a method for pre-training language models on large amounts of unlabeled text. This method, known as BERT, has been widely adopted in natural language processing (NLP) and has improved performance on a variety of downstream tasks. Lample and Conneau [81] proposed a method for unsupervised machine translation, in which a model is trained to translate between languages without using any parallel data. The method is based on adversarial training and denoising autoencoding, and has shown promising results on several low-resource language pairs.

Gu et al. [82] proposed a meta-learning approach for unsupervised neural machine translation, which was shown to be effective for several low-resource language pairs. Artetxe et al. [24] also proposed a method for unsupervised neural machine translation with weight sharing, which was also shown to be effective for low-resource language pairs. The approach involves sharing the encoder and decoder weights across multiple languages, allowing the model to leverage knowledge from other languages to improve translation quality. Lample et al. [83] proposed a multilingual transformer model that can be trained on multiple languages simultaneously, which was shown to improve the performance of low-resource machine translation systems. This approach involves sharing the parameters of the model across multiple languages, allowing the model to leverage the shared information to improve translation quality. Adversarial training, meta-learning, and weight sharing across different languages are indeed promising approaches for addressing the challenges of low-resource languages. However, there are two notable limitations associated with these methods. The first limitation is their computational expense. Adversarial training and meta-learning techniques often require significant computational resources, including high-performance hardware and extensive training time. This can hinder their practicality and scalability, particularly in resource-constrained settings where computational resources are limited. The second limitation arises from the difficulty of sharing weights across languages with different linguistic structures. Languages with distinct morphological characteristics, such as Afaan Oromo, present additional challenges in weight sharing. This makes it challenging to directly transfer weights or apply shared models across languages with diverse structures.

Given these limitations, Lample and Conneau [25] proposed a method for unsupervised learning for neural machine translation using monolingual data only. The method involves pre-training

a language model on a large amount of monolingual data and then fine-tuning the model on parallel corpora. The approach was shown to be effective for several low-resource language pairs, including English-Afaan Oromo. But, this approach also has its own limitation which we are going to discuss below.

Upon thorough evaluation of the aforementioned content and additional relevant research, we can categorize the concepts pertaining to the utilization of PTLMs for enhancing MT through sequence-to-sequence language models into three distinct classes. These classes encompass the strategies of Initialization and Fine-tuning, employing PTLMs as Context-aware Embeddings, and adopting Teacher-Forcing and Knowledge Distillation approaches.

Initialization and Fine-tuning: One approach is to initialize the parameters of downstream models with PTLMs and then fine-tune them on specific tasks. This strategy, proposed in the influential BERT paper by Devlin et al. [16], has been widely adopted in various natural language processing tasks. However, Zhu et al. [26] experimented with this approach for MT tasks and found no significant improvement, even when using models like XLM [25] with large parallel corpora. This suggests that while initialization and fine-tuning may be effective in some scenarios, its impact on MT tasks with under-resourced languages might be limited.

Context-aware Embeddings: Another strategy is to use PTLMs as context-aware embeddings for downstream tasks. This approach, inspired by the work of Peters et al. [27], aims to enhance translation quality by incorporating contextual information from PTLMs. Zhu et al. [26] further explored this method and proposed a more promising approach called the BERT-Fused approach. By leveraging the contextual embeddings provided by PTLMs, this strategy showed potential for improving translation quality in a more effective manner.

Table 4. *Preliminary explorations on IWSLT'14 English→German translation [26]*

Algorithm	**BLEU Score**
Transformer	28.57
Initialize the encoder of NMT using BERT	27.14
Initialize the encoder of NMT using XLM	28.22
Initialize the decoder of NMT using XLM	26.13
Initialize both the encoder and decoder of NMT using XLM	28.99
Leveraging the output of BERT as embeddings	29.67

From table 4 experimented by Zhu et al. [26], we can see that by leveraging the output of BERT as embeddings the BLEU score obtained was 29.67, which is the highest among all the approaches. This indicates that leveraging the contextual embeddings from BERT led to the

most significant improvement in translation quality compared to the original transformer model. Using this concept as the path of the lighter, Zhu et al. [26] have shown the way we can fuse the output of BERT to the original transformer model by introducing new cross attention between BERT's output and encoders/embeddings output in each layer.

Teacher-Forcing and Knowledge Distillation: The third strategy involves using PTLMs as a teacher forcing method in knowledge distillation (KD) approaches. While the two previous strategies mainly focus on language understanding tasks, Chen et al. [28] studied the use of PTLMs for text generation and distilled their knowledge for downstream tasks. They presented a novel approach, Conditional Masked Language Modeling (C-MLM), to finetune BERT on language generation. Using the unique bidirectional characteristics of BERT, the process of distilling the acquired knowledge in BERT can motivate auto-regressive Seq2Seq models to anticipate future steps, thereby applying comprehensive sequence-level guidance to ensure coherent text generation. This approach can be beneficial in improving the quality of generated text, such as in machine translation or other language generation tasks.

Based on the literature mentioned, again the concepts can be divided into two distinct categories: language understanding and language generation. Previous researches have predominantly focused on either one of these categories, with limited exploration of simultaneously considering both aspects in improving translation quality for under-resourced languages. This presents an opportunity for future studies to investigate approaches that integrate language understanding and generation to address the specific challenges faced by under-resourced languages in machine translation. By developing novel techniques that leverage PTLMs in a holistic manner, researchers can potentially achieve significant advancements in the translation quality for these languages.

Let's discuss further the incorporation of PTLMs into NMT systems, focusing on the aspects of language understanding and language generation.

Language Understanding in NMT: Language understanding in NMT refers to the ability of the system to comprehend the source language and accurately capture its semantics before generating the target translation. PTLMs can be leveraged to enhance language understanding in NMT through the following approaches: Initialization and Fine-tuning and Context-aware Embeddings.

As mentioned earlier, one strategy is to initialize the encoder model of the NMT system with PTLMs and then fine-tune it on the specific translation task. This initialization can provide the NMT system with a strong language representation foundation, enabling it to better understand the source language. Fine-tuning allows the system to adapt the PTLM's general language

understanding capabilities to the specific translation task, potentially improving the quality of translations for under-resourced languages. PTLMs can also be used as context-aware embeddings in NMT. Instead of initializing the entire model, the source language sequence can be encoded using PTLMs to obtain contextualized representations. These representations capture the meaning and context of the source text, aiding the NMT system in better understanding the input and generating more accurate translations. This approach allows for the integration of PTLMs' contextual information without the need for extensive fine-tuning.

Language Generation in NMT: Language generation in NMT refers to the process of generating fluent and coherent target translations based on the comprehension of the source language. PTLMs can contribute to improving language generation in NMT through the following Teacher-Forcing and Knowledge Distillation technique. The PTLM can act as a teacher model in a knowledge distillation framework, guiding the generation process of the NMT system. The PTLM's rich language generation capabilities and knowledge can be transferred to the NMT system, promoting better fluency, coherence, and overall translation quality. By training the NMT system to mimic the PTLM's generation patterns, it can produce more natural and contextually appropriate translations for under-resourced languages.

In conclusion, unsupervised pre-training methods such as BERT [16] and GPT [29] have shown remarkable success in improving the performance of NLP tasks by training large-scale language models on massive amounts of unlabeled data. These models can learn general-purpose language representations that capture complex linguistic structures and patterns, which can then be fine-tuned for specific downstream tasks with limited labeled data. Similarly, pre-training transformer models on large amounts of monolingual data from the low-resource language, followed by incorporating them to the original transformer model and training on a smaller amount of parallel data, can potentially improve the performance of NMT for that language pair. This approach can help the NMT model learn better representations of the low-resource language, which can be particularly important for languages with complex morphologies and syntax. However, these pre-training language models solely depend on either of Unsupervised Language Understanding (LU) or Language Generation (LG) tasks, i.e., they are not strong in both LU and LG at the same time. This will directly affect the translation quality by lacking consideration of both source side and target side languages together. Therefore, a method which combines both LU and LG could enhance the translation quality. But, to the level of our knowledge there is no work which applied these approaches on the original Transformer model by marrying the two concepts of language understanding and language generation. In this study, we aspire to tackle these challenges by incorporating BERT for LU and GPT for LG to the transformer model and applying a monolingual corpus based unsupervised pre-training method.

Chapter Three

3. Methodology

3.1 Introduction

The success of any NLP system largely depends on the quality and quantity of data used to train it. In this study, we aim to develop an English-Afaan Oromo machine translation system by leveraging the power of unsupervised pre-training and incorporating of state-of-the-art language models into Transformer model. To achieve this, we collected a large monolingual unlabeled text corpus for Afan Oromo, which is used to fine-tune both the pre-traied BERT and GPT models. We then incorporated these pre-trained models on a parallel English-Afaan Oromo dataset to obtain the final translation system to transformer model by using additional attention mechanisms. In this section, we will describe in detail our methodology for data collection, fine-tuning, Incorporating, and evaluation of the proposed machine translation system.

3.2 Dataset

The data collection section is a crucial step in any natural language processing (NLP) project. In this study, the aim is to fine-tune pre-trained language models using unsupervised pre-training method on monolingual data in Afan Oromo, an under-resourced language with a limited availability of labeled data. Therefore, the data collection process is critical to ensure the success of the thesis.

The first step in this study was to collect a monolingual text corpus of Afan Oromo, which is used for unsupervised pre-training. To collect the data, we used both primary and secondary data sources. We used publicly available text sources from BBC Afaan Oromoo's website, Fana Brodcasting Corporation's website, books, and other useful websites to compile the corpus as primary sources. Secondly, we used the Afan Oromo text corpus as secondary source collected by Ogueji et al. [84] on their AfriBERTa paper which is freely available on huggingface[1]. These sources are explored to identify any text in Afan Oromo that can be used for training purposes.

Additionally, we used the parallel English-Afaan Oromo dataset to evaluate our proposed approach on Transformer model. The objective of this study is to handle the lack of large parallel dataset for effective machine translation. Therefore we used the parallel dataset used by Waldeyes [33], the dataset collected by Asmelash Teka which is collected by HornMT[2] project and publicly available on Github, and we collected ourselves from different sources including

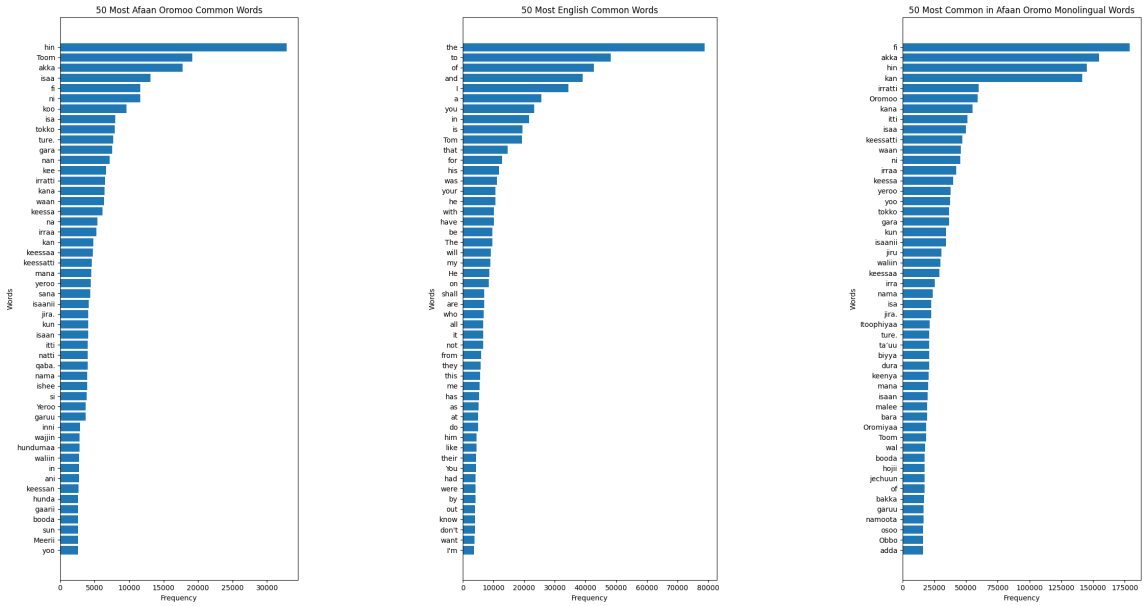
[1]available at https://huggingface.co/datasets/castorini/afriberta-corpus
[2]available at https://github.com/asmelashteka/HornMT

Holly Bible to measure the effectiveness of our approach. Table 5 illustrates the sets of parallel and monolingual data utilized in this research, along with their respective distributions based on the size of parallel and monolingual sentences, unique words, tokens, and vocabulary size.

Table 5. *English-Afaan Oromoo parallel and Afaan Oromoo monolingual dataset distribution.*

Languages	**Sentences**	**Tokens**	**Unique Words**	**Vocabulary Size**
Afaan Oromoo	84,565	519,552	40,964	23,844
English	84,565	631,839	33,534	15,518
Afaan Oromoo monolingual	712,358	10,648,319	748,500	47,348

Our final dataset includes 712,358 monolingual Afaan Oromo sentences and 84,565 parallel sentences for the English-Afaan Oromo language pair, as Table 5 illustrates. There are 631,839 tokens and 33,534 unique terms in English, 519,552 tokens and 40,964 unique words in Afaan Oromo of the English - Afaan Oromo parallel dataset. Similiarly, there are 10,648,319 tokens and 748,500 unique words in the Afaan Oromo 712,358 monolingual sentences. In the parallel dataset, Afaan Oromo contains less tokens and more unique words than English. This is because the Afaan Oromo language has morphological inflection and derivation, which results in a great number of variations for a single word.

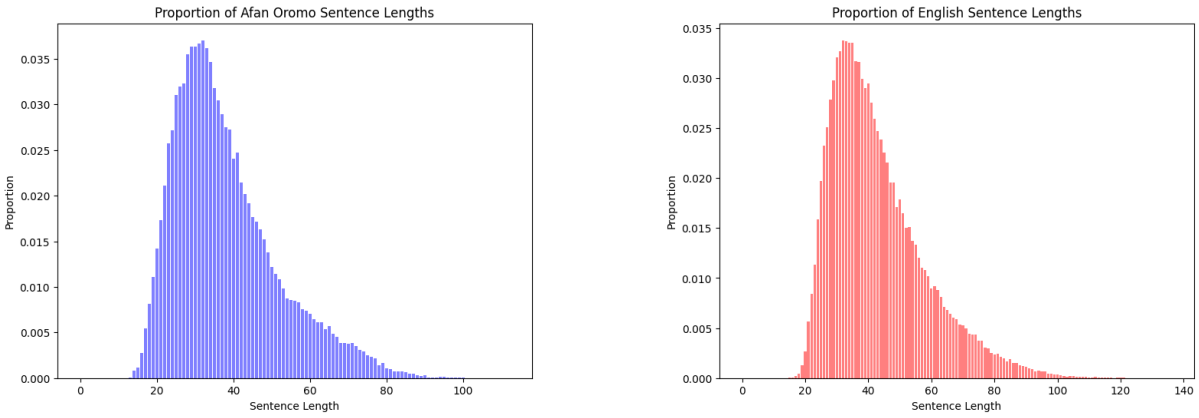


(a) (b) (c)

Figure 2. *Fifty most common words in the English-Afaan Oromo parallel corpora and Afaan Oromo Monolingual dataset. 2a Afaan Oromo, 2b English, and 2c Afaan Oromo Monolingual Words.*

In a similar way, when comparing the monolingual dataset with the Afaan Oromo data in parallel dataset, the latter has big difference in words. From 712,358 sentence corpus, we get around 10M tokens. This can help us to get more unique words to introduce for Transformer model. This creates great opportunity for the Transformer model to learn more grammatically complex and morphological rich representations that can boost the performance of the model in translation process.

Figure 2 shows the fifty most common tokens in the parallel and monolingual datasets, respectively. Figures 2a and 2c illustrate how the most prevalent words across all domains are shared by the parallel and monolingual Afaan Oromo data. The Afaan Oromo monolingual data covers the sports, politics, news, education, and more domains, while the English-Afaan Oromo parallel dataset is primarily in the religion area. For example, words such as hin (not), fi (and), akka (like), nama (man), hundumaa (all), gaarii (good) are appear in both datasets, but words such as, Oromoo (Oromo), Itoophiyaa (Ethiopia), biyya (country), bara (year), Obbo (mister) are appiered only in monolingual data. This helps the model to generalize the context by getting additional insights from monolingual data.



(a) (b)

Figure 3. *Sentence length and Density of Parallel Dataset 3a Afaan Oromo 3b English.*

Figure 3 depicts the length of sentences for both languages by their proportion. By observing both figure 3a and 3b, we can see both languages of parallel dataset most sentences are between the sentence length of 3 - 100. The maximum sentence length for Afaan Oromo is 100 and 120 for English.

Figure 4 demonstrates the contrast in sentence lengths between English and Afan Oromo data. According to the figure, it is evident that Afan Oromo sentences tend to be shorter compared to English sentences. Due to the morphological complexity of Afaan Oromo, translating a sentence into Afan Oromo often requires fewer words. For instance, the sentence "Turn off the light

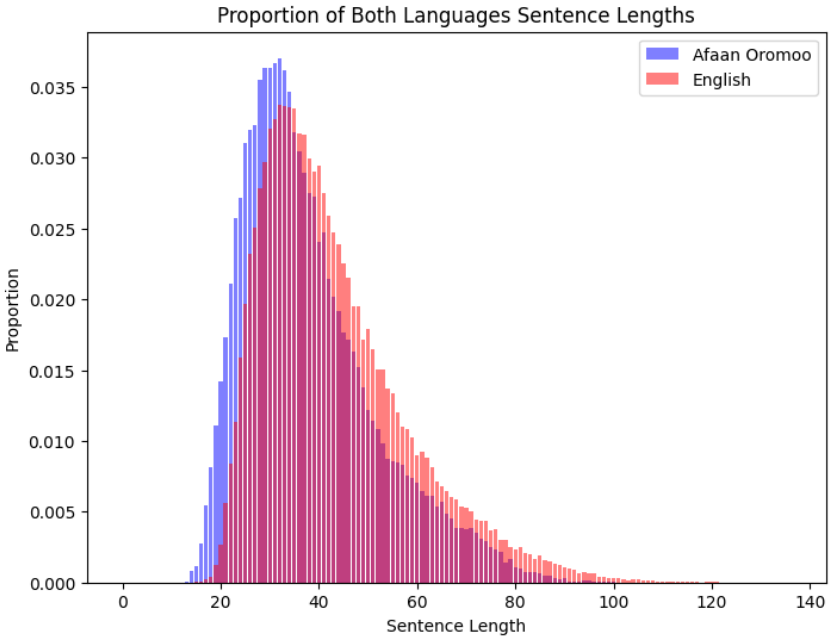


Figure 4. *Sentence pair proportion for sentences up to a maximum length of 120.*

and go to sleep." translates to "Ifaa dhaamsii rafi." in Afan Oromo. While the English sentence consists of eight words, the Afan Oromo translation only uses three words to convey the same meaning. As the number of words in a sentence increases in Afan Oromo, English sentences begin to utilize more words than their Afan Oromo counterparts.

Index	Afaan Oromo	English
0	Haasaan isaa ayyaanichaaf kan mijatu ture.	His speech was suitable for the occasion.
1	Gargaarsi na barbaachisa jette.	You said you needed help.
2	Vaayirasichi jijjiiramuu eegaleera.	The virus is starting to mutate.
3	Muuxannoo hadhaawaa isa mudate.	He had a bitter experience.
4	Hiriyaan koo fi abbaan warraa ishee lamaan isaanii iyyuu Ameerikaanota dha.	My colleague and her husband are both American.
5	Asitti dhukkubbii itti fufiinsa qabun qaba.	I have a persistent pain here.
6	Konkolaattota biyya alaatii galfaman irratti gibirri addaa kaa'ameera.	A special tax was imposed on imported cars.
7	Nu waliin dhufta jedheen akka waan salphaatti fudhadhe.	I took it for granted that you would come with us.
8	Toom asoosama gaggabaaboo hedduu barreessee jira, garuu hanga ammaatti kitaaba hin maxxansine.	Tom has written several short stories, but hasn't yet published a book.
9	Inni nama baayyee walii hin galledha.	He's a very disagreeable man.

Figure 5. *Sample random texts from English-Afaan Oromo Parallel Dataset*

Figures 5 and 6 excerpts the screenshoot of few sample texts from the parallel and monolingual datasets. Screenshoot on Figure 5 shows us the parallel dataset. The left side sentences are Afaan Oromo sentences and the right side sentences are English sentences.

Table 6 presents the translated sample text for the Afaan Oromo monolingual dataset displayed in Figure 6, which was manually translated from the monolingual source.

Index	Afaan Oromo Monolingual Data
0	Filannoo kana geeggessuuf rakkoo maallaqaa fi meeshaalee barbaachisoo ta'an dhiheessuun cinatti filannicha taasisuuf yeroon jiru ji'a tokko kan hin caalle ta'uusaatiin qophiilee dursanii taasisuun rakkisaa waan ta'uuf filannicha yeroo gabaabaa kanatti taasisuun rakkisaa ta'uu mala jedhamaa ture.
1	Haaluma kanaan har'as Harargee magaaloto akka Awwadaay, Baddeessaa, Dadara, Hirnaa, Qobboo, Haramaayaa fi iddoo addaddatti mormii fi dallansuun ummataa cimatu dhaggeesifamaa oole.
2	Ummatnisi kana hubatee akka qoratamuufi yoo akka tasaa vaayrasichaan qabames gorsa ogeessonni fayyaa kennaniin bakka wal'aansaa qophaa'etti tajaajila argachuun akka danda'amus eeran.
3	Biig Maapil Liif, warqeen jajjabaan saantima doolaaraa Kanaadaa miliyoona 1 tahu kan kiilogriraama 100 (220 lb) ulfaatu, Godambaa Bood kan Beerlin, Jarman keessaa hatame.
4	"Waldaan Waaqayyoo hundinuu akkana jedha, 'Isin har'a Yihowaa duukaa bu'uu irraa garagalchuuf, har'a Waaqayyoon irratti fincila kaasuudhaaf, iddoo aarsaa ofii keessaniif ijaartanii, Waaqayyo Israa'el irratti yakka maalii dha?
5	Maricharratti Preezdaantii Mootummaa Naannoo Oromiyaa obbo Shimallis Abdiisaa fi Ministira Ministeera Tuurizimii Ambaasaaddar Naasisee Caalii dabalatee hoggantoonni Abbaa Taayitaa Eegumsa Bineensota Bosonaa Ityoophiyaa fi Komishinii Tuurizimii Oromiyaa hirmaataniiru.
6	Toom kitaabota Faransaayii hedduu dubbiseera.
7	Diyaamomdiin halluu burtukaanaaa abbaan karaatii 14.82 erga caalbaasii Kiristii kan Jeneevaa keessaa irratti miliyoona $35.5'n gurguramee rikerdii addunyaa qabate.
8	Phaawuloos warra Roomaatiin yeroo barreessu bara Neerootti harka warra Roomaa ture namni kun Roomaa kan turan bitootni hunduu kan haammate ture.
9	Ofiis nama waraanaa ta'uuf yaada qabu turan.

Figure 6. *Sample random texts from Afaan Oromo Monolingual Dataset*

Table 6. *Manually translated sample text for Afaan Oromo monolingual text shown on Figure 6*

Afaan Oromo Monolingual Equivalent Translation (Manually Translated)
It was said that it would be difficult to hold the election in such a short period of time as it would be difficult to make preparations in advance as the time left for the election is less than a month.
There are many towns in Hararghe today suchas Awaday, Bedessa, Dadara, Hirna, Haramaya adn other places where protests and anger were heard.
He said the people should be aware of this and if they are accidentally infected with the virus, they can get services at the treatment facilities provided by health professionals.
Big Maple Leaf, a 100-kilogram (220 lb) solid gold coin worth 1 million Canadian dollars, was stolen from the Bod Museum in Berlin, Germany.
The whole congregation of God says, 'What is this guilt against the God of Israel, that you have built yourselves an altar this day to turn away from following the Lord, to rebel against God this day?
The discussion was attended by the President of the Oromia Regional State, Mr. Shimelis Abdissa and the Minister of Tourism, Ambassador Nasise Chali.
Tom has read many French books.
A 14.82-carat orange diamond set a world record after selling for $35.5 million at Christie's in Geneva.
When Paul wrote to the Romans, he was in the hands of the Romans during the reign of Nero.
They intended to become warriors themselves.

3.3 Pre-processing of dataset

In this section, we describe the data pre-processing steps for our proposed approach to English-Afaan Oromo machine translation, which involves fine-tuning of language models on a large unlabeled monolingual dataset, followed by incorporating the language models into transformer model on a parallel English-Afaan Oromo dataset. The quality of the final machine translation system heavily depends on the quality of the input data and how it is pre-processed. Pre-processing involves cleaning the text dataset to prepare it for inputting into a model. This step is crucial to eliminate unwanted characters and words from the dataset. The dataset often contains noise in the form of punctuation and text written in different cases and so forth. Therefore, in this section, we explain the data pre-processing steps we take to ensure the highest possible quality of the training data for both the fine-tuning and incorporating stages of our approach.

The first step in our approach is to fine-tune the language models using a large unlabelled monolingual dataset of Afaan Oromo. We preprocessed the dataset in the following way:

Lowercasing: Lowercasing is a common step in data pre-processing for various NLP tasks. All words are converted to lowercase to reduce the dimensionality of the vocabulary. It involves converting all the text in a given dataset to lowercase letters. Lowercasing is performed for several reasons. Lowercasing helps to normalize the text by reducing the number of distinct word forms. By converting all text to lowercase, variations in capitalization (e.g., "Akkam" vs. "akkam") are removed, treating them as the same word. This normalization step ensures that words with the same spelling but different capitalization are treated consistently, preventing the model from learning separate representations for them. Lowercasing reduces the size of the vocabulary. In many cases, words that differ only in capitalization have similar meanings and can be considered as the same word for the purpose of NLP tasks. By converting everything to lowercase, the vocabulary size is reduced, leading to more efficient memory usage and faster computation during training and inference. Lowercasing helps the model generalize better by treating different capitalizations of the same word as equivalent. This is particularly useful when dealing with out-of-vocabulary (OOV) words or rare words. By converting them to lowercase, the model has a higher chance of encountering similar words during training and can learn more robust representations. Therefore we applied lowercasing step while pre-processing step.

Cleaning: We remove any special characters, URLs, email addresses, and other non-alphabetic symbols from the text. Removing special characters, URLs, email addresses, and other non-alphabetic symbols from text is an important step in data pre-processing for several reasons. Special characters, URLs, email addresses, and non-alphabetic symbols often introduce unnecessary noise to the text data. They can disrupt the analysis or modeling process by adding irrelevant information or patterns that might not be meaningful for the specific task at hand.

Removing these elements helps to clean the text and focus on the essential content. Special characters and non-alphabetic symbols can make the text more difficult to process or analyze. Removing them simplifies the text and facilitates subsequent processing steps. By removing them, the vocabulary size is reduced, which can lead to more efficient memory usage and faster computation during training and inference. This reduction in vocabulary size can also help mitigate the data sparsity problem.

In this step we kept the apostrophe (') special character as it is. Because apostrophe in Afaan Oromoo writing system is considered as one character. It refers to the hudhaa sound, which is heard in the words "re'ee" and "fal'aana" as cited on paper Tegegne [32]. More detail about apostrophe in Afaan Oromoo is dicussed under Section 2.1.1.1

Again we don't remove some special characters such as "!", "?" and ".". Preserving these punctuation marks during text pre-processing for machine translation is recommended for several reasons. Firstly, these punctuation marks contribute to the structure of a sentence. They indicate the end of a sentence (period), a question (question mark), or an exclamation (exclamation mark). By retaining these marks, the machine translation system can better understand the intended sentence structure, resulting in more accurate translations. Secondly, punctuation marks provide valuable contextual cues that help disambiguate the meaning of a sentence. A question mark at the end of a sentence indicates a question, while an exclamation mark denotes emphasis or strong emotion. By preserving these marks, the machine translation system can capture the intended tone, intention, or mood of the original sentence, leading to more precise translations. Additionally, preserving punctuation marks facilitates post-processing and evaluation of the translated text. If punctuation marks are removed during pre-processing, it becomes challenging to correctly re-insert them in the translated output. Moreover, punctuation marks are often considered during evaluation metrics for machine translation, and their presence can impact the evaluation scores. But for consistency and to get them separately while tokenization, we added extra space between other characters and these special characters.

We then strip whitespace to remove extra spaces from the sentence. Stripping whitespace is a common step in text pre-processing that involves removing leading and trailing spaces or other whitespace characters from a text. This process serves several purposes, including removing extraneous characters, consistent tokenization, and normalizing text. Whitespace characters like spaces, tabs, or line breaks can be considered noise in the text data. By stripping whitespace, we eliminate unnecessary characters that do not contribute to the actual content or meaning of the text. Stripping whitespace helps ensure consistent tokenization, especially when using whitespace as a delimiter. By removing leading and trailing spaces, we prevent tokens from being mistakenly split or merged due to inconsistent whitespace usage. In our parallel dataset, we use tab to separate source sentence from target sentence. Stripping whitespace is part of the

text normalization process, where we aim to standardize the representation of text. Normalizing whitespace ensures that multiple consecutive spaces or other whitespace characters are reduced to a single space, promoting uniformity in the text.

Filtering: We remove any sentences that are shorter than 3 words or longer than 120 words to remove noise from the dataset. Removing sentences that are shorter than 3 words or longer than 120 words is a common practice in data pre-processing to eliminate noise from the dataset. Short sentences are often incomplete or lack sufficient context, making them less informative for training a model. On the other hand, excessively long sentences may contain complex structures, repetitions, or irrelevant information that can hinder the learning process or introduce noise into the model. The original Transformer model introduced by Vaswani et al. [46] takes the maximum length of sentence containing 128 terms. By setting these length thresholds, we can filter out sentences that are likely to be uninformative or noisy, allowing the model to focus on more meaningful and representative data. This filtering step helps to improve the quality and relevance of the dataset, leading to better performance and generalization of the model. Additionally, removing extremely short or long sentences can help to ensure consistency in the length of input sequences, which is often desirable for efficient processing and modeling.

Tokenization: The raw text is split into individual words using the whitespace as a delimiter. We use the tokenization method provided by the Hugging Face Tokenizers library. Utilizing the tokenization method provided by the Hugging Face Tokenizers library to split raw text into individual words using whitespace as a delimiter offers several benefits, particularly in maintaining consistency with BERT and GPT models.

Firstly, consistency in tokenization is essential for achieving accurate and reliable results when working with pre-trained language models like BERT and GPT. These models are trained on specific tokenization schemes, and using a different tokenization approach can lead to incongruities between the training and inference phases. By employing the Hugging Face Tokenizers library, which aligns with the tokenization scheme employed during fine-tuning, we can ensure that the input data is processed in a manner consistent with the models' expectations. This consistency enhances the quality of downstream tasks, such as machine translation, as the models can effectively understand and interpret the tokenized input.

Another reason to use Hugging Face Tokenizers library in this context is its ability to handle specialized cases and language-specific tokenization requirements. Different languages may have unique linguistic structures, such as compound words or morphologically rich forms, that necessitate specific tokenization approaches. The Hugging Face Tokenizers library offers various tokenization algorithms and pre-trained models designed to handle specific languages, enabling accurate and linguistically-aware tokenization. This language-specific tokenization

ensures proper handling of language-specific nuances and contributes to improved performance and understanding in multilingual or cross-lingual contexts. Additionally, the Hugging Face Tokenizers library provides efficient and optimized tokenization implementations, enabling faster processing of large volumes of text.

The second dataset used in our approach is a parallel corpus of English and Afaan Oromo sentences. This dataset is preprocessed with the same as step monolingual corpus discussed above except:

Alignment: We align the sentences in the parallel corpus at the sentence level separated by tab.

After pre-processing, the monolingual dataset is saved as a plain text file with one sentence per line. This pre-processed dataset is then used to fine-tune the BERT and GPT models. The parallel corpus is saved in the standard format of the fairseq toolkit for training the transformer model.

3.4 System Architecture

The proposed system architecture for English-Afaan Oromo machine translation includes both the pre-processing and post-processing stages. The pre-processing stage includes data collection and data pre-processing, while the post-processing stage includes the actual machine translation process. The system architecture is based on the Transformer model, which is a state-of-the-art model for machine translation. The Transformer model is composed of an encoder and a decoder, each containing a stack of attention layers. The encoder takes the source language sentence as input and produces a sequence of hidden states, while the decoder takes the hidden states and the previously generated target language tokens as input and generates the next target token.

The pre-trained language models used in this work are the BERT and GPT models. The models are first fine-tuned on monolingual text corpus, and then the source text is given for the pretrained language model for the machine translation task. The incorporating process is done by training the entire model end-to-end on parallel English-Afaan Oromo corpus with a sequence-to-sequence objective function. Figure 7 shows the whole system architecture for this paper.

3.4.1 Baseline

In order to evaluate the effectiveness of our proposed method in improving machine translation quality for English-Afaan Oromo, it is important to establish a baseline for comparison. Our

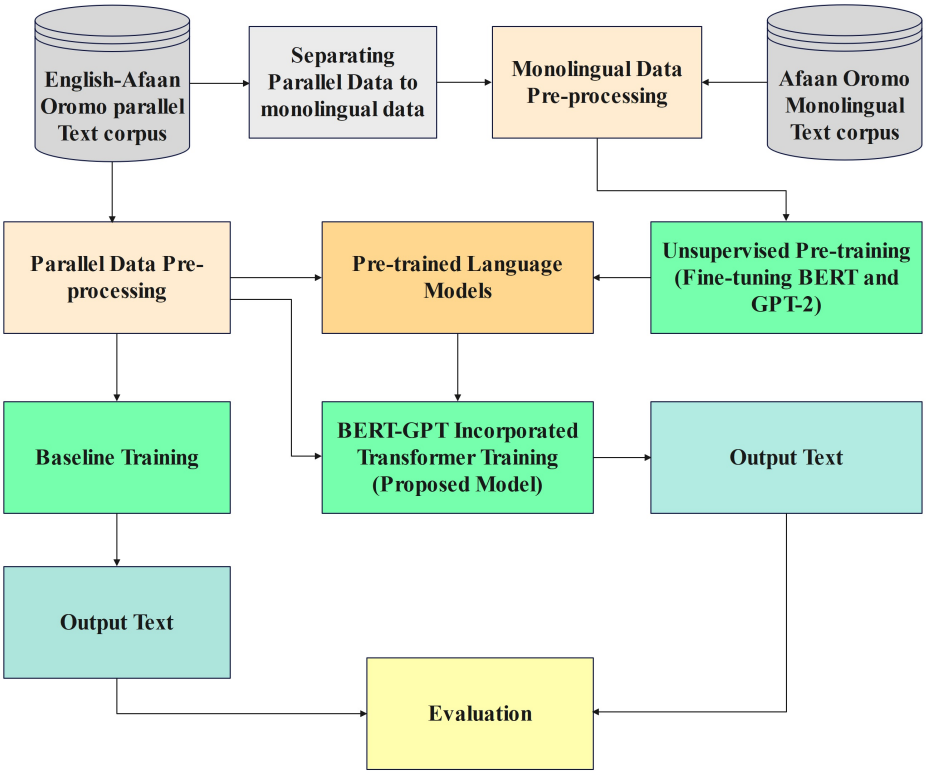


Figure 7. *System Architecture*

proposed approach utilizes a Transformer-based model, so the first baseline we consider is the original Transformer model as described in the paper by Vaswani et al. [46]. This serves as a benchmark that we can measure our improvements against.

Additionally, we also include a second baseline approach known as the BERT-Fused approach, which was proposed by Zhu et al. [26]. This allows us to assess the performance of our method relative to other unsupervised approaches that incorporate pre-training mechanisms.

3.4.1.1 Transformer Model

In the context of seq2seq problems like NMT, the initial approaches relied on using RNNs within an encoder-decoder architecture [52, 53, 85]. However, these architectures faced limitations when dealing with long sequences [54]. As new elements were added to the sequence, the encoder's ability to retain information from the initial elements diminished. The hidden state in each step of the encoder was associated with a specific word in the input sentence, typically one of the most recent words. Consequently, if the decoder only had access to the last hidden state of the encoder, it would lack important information about the earlier elements of the sequence. To address this limitation, the attention mechanism was introduced [46].

Instead of focusing solely on the last state of the encoder, as is conventionally done with RNNs, attention allows each step of the decoder to consider all the states of the encoder [46]. This enables access to information about all the elements in the input sequence. The attention mechanism extracts information from the entire sequence by calculating a weighted sum of the past encoder states. This empowers the decoder to assign greater importance to specific elements of the input when predicting the next output element. During the learning process, the model learns to dynamically focus on the relevant input element for each output element. And the model is designed to be trained in parallel way, so that we can train it on GPU accelerator. This transformer achieved SOTA in Machine Translation. Therefore we selected Transformer model to compare with and to see the effectiveness of our proposed model.

3.4.1.2 BERT-Fused Model

Incorporating the pre-trained language models to the original Transformer model is proposed by Zhu et al. [26] and achieved SOTA benchmark over other previously proposed unsupervised approaches and our proposed paper is inspired by this paper. The paper introduced us how to incorporate the BERT output to the original transformer model by introducing additional attention wich are BERT-Encoder attention and BERT-Decoder attention. Therefore we have chosen this approach as a baseline.

By comparing our proposed approach against these two baselines, we can gain insights into the extent to which our method surpasses existing techniques and enhances the quality of machine translation.

3.4.2 Unsupervised Pre-training (Fine-tuning)

Unsupervised pre-training is a popular method used to improve the performance of natural language processing (NLP) models. The idea is to train a model on a large amount of unlabeled data, typically using a self-supervised learning approach, and then use the model on a downstream task using labeled data. This can be a highly effective way to improve the performance of NLP models, especially for under-resourced languages where labeled data is of limited availability.

The pre-training phase involves fine-tuning the language models on the monolingual corpus using BERT and GPT. In our work we fine-tune the pre-trained language models for our downstream MT task rather than pre-training them from scratch. Fine-tuning is a common approach used in machine learning when working with pre-trained models. Instead of training a model from scratch, fine-tuning allows us to leverage the knowledge and expertise captured by a pre-trained model and adapt it to a specific task or domain.

One advantage of fine-tuning pre-trained language models is that they can learn general language representations that can be applied to a wide range of downstream tasks. This can help to overcome the problem of limited labeled data for under-resourced languages, as the pre-trained models can capture a lot of the underlying structure and patterns in the language.

As previouly discussed, the Transformer model has achieved many SOTA benchmarks in different NLP tasks. From these MT is the one. But this transformer model needs more parallel data to learn complex language representation and to generate high quality target language. The transformer model is consisting of Encoder and Decoder methods. The encoder is to learn the representation of source language and the Decoder is to learn generation text in target language.

To solve this we have to focus on both language representation (language understanding) and language generation tasks. For language understanding, BERT achieved better result [16]. Because BERT is the Encoder based approach. Therefore we have selected BERT model to learn the general representation of source language and the last layer output of BERT's model is incorporated to the original model. So that the model can simply learn the representation of a source sentence. We use the Pre-trained BERT LM and fine-tune it on our available English sentences on parallel dataset for English sentences when English is considered as source language and we use the pre-trained BERT LM and fine-tune it on collected monolingual data for Afaan Oromo sentences when Afaan Oromo is considered as source sentence. The BERT model is trained in the fashion of predicting masked word in a sentence. This approach is called Masked Language Model (MLM). In MLM the model can see the words in both left and right side then predicts the masked word in simple way. This is called Bi-Directional.

For language generation, GPT-2 achieved better result [29]. GPT-2 is Decoder based approach. It has better performance in generating next word in a given sequence. By learning sentence generation in particular language, GPT-2 can act as Teacher model and transformer model as student by using KD approach. Therefore, we have selected the pre-trained GPT-2 model to fine-tune it on available sentences from parallel dataset for generating next word in the same vocabulary with transformer model. This helps the model to learn generating high quality text in limited parallel corpora.

3.4.3 Incorporating (Proposed Approach)

Incorporating is the process of fusing pre-trained language models with a Transformer model. In a paper written by Zhu et al. [26], it is suggested that it is beneficial to incorporate pre-trained language models into the Transformer model, rather than simply fine-tuning them or using them solely as encoders.

Building upon Zhu's architecture, we propose a modified approach that leverages the strengths of both BERT and GPT models. Our goal is to enhance the performance of the original Transformer model by utilizing BERT as a context-aware embedding and incorporating it into both the encoder and decoder components. Simultaneously, we employ GPT for language generation purposes and utilize it as a "teacher model" to guide the decoder in predicting the next word by using Knowledge Distillation(KD) approach.

Notations: Let X and Y represent the domains of the source and target languages, respectively. These domains consist of collections of sentences in their respective languages (English and Afaan Oromo in this case). For any given sentence x in X and sentence y in Y , we denote the number of units (such as words or sub-words) in x and y as n_x and n_y, respectively. The i-th unit in x or y is represented as x_i or y_i. In our work, we refer to the encoder and decoder as the NMT module for simplicity. Without loss of generality, we assume that both the encoder and decoder have N layers. The attention layer, denoted as $attn(q, K, V)$, is used in our model, where q, K, and V represent the query, key, and value, respectively, following the notation used in Vaswani et al. [46]. We utilize the same feed-forward layer as described in Vaswani et al. [46] and denote it as FFN.

3.4.3.1 BERT-Fusion

Specifically, we fine-tune the pre-trained BERT model on a monolingual dataset for English-Afaan Oromo. The output of BERT, which provides contextualized word representations, will be incorporated into both the encoder and decoder layers of the Transformer model. This incorporation will enable the model to benefit from the contextual information captured by BERT.

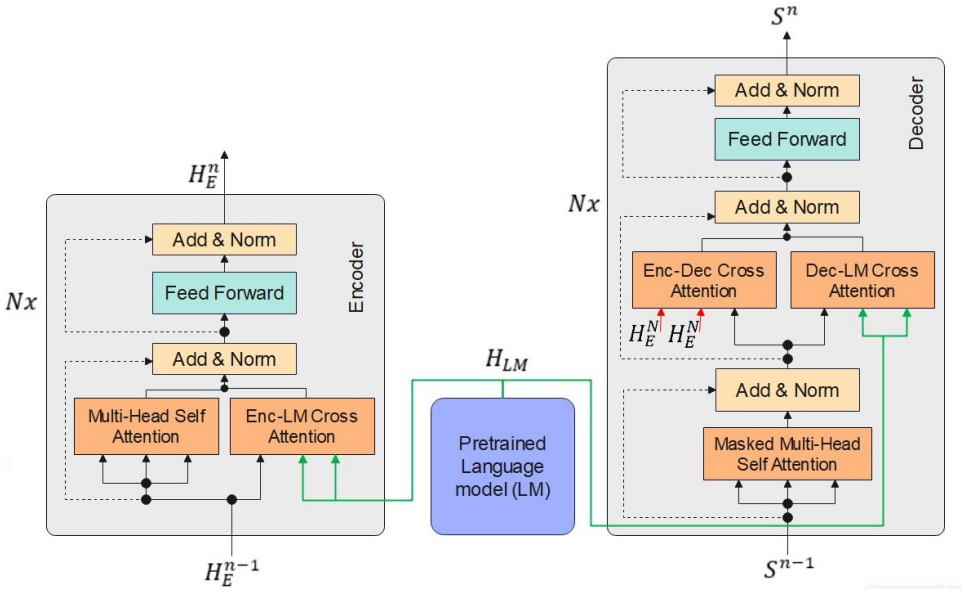


Figure 8. *The BERT-Fusion Model as Proposed by Zhu et al. [26]*

Figure 8 illustrates the suggested architecture for incorporating BERT into a transformer model. In the figure, the encoder, decoder, and pre-trained Language Model (BERT) are denoted as Enc, Dec, and LM respectively. The output of the Language Model (BERT) is represented by the green lines (H_{LM}). Similarly, the output of the final layer of the encoder is indicated by the red lines (H_E^N). The dashed lines indicate the residual connection.

The output of BERT is incorporated in the following steps:

Step-1: For any input x from the set X, LM takes x and encodes it to obtain the representation H_{LM}.

$$H_{LM} = LM(x) \tag{3.1}$$

H_{LM} is the output of the final layer in LM. Each $h_{LM,i} \in H_{LM}$ represents the i-th wordpiece in x.

Step-2: Let denote the hidden representation of the n-th layer in the encoder as H_E^n. Additionally, the word embedding of the sequence x is denoted as H_E^0 and the hidden representation of the last layer of encoder is denoted as H_E^N. Let the i-th element in H_E^n is represented as h_i^n for any $i \in [n_x]$. In the n-th layer of the encoder, $n \in [N]$,

$$\tilde{h}_i^n = \frac{1}{2}(attn_S(h_i^{n-1}, H_E^{n-1}, H_E^{n-1}) + attn_{LM,E}(h_i^{n-1}, H_{LM}, H_{LM})), \forall i \in [n_x] \tag{3.2}$$

In this concept, we have two attention models called $attn_S$ and $attn_{LM,E}$, each with its own set of parameters (defined in Eqn. 3.6). These attention models are applied to a sequence of hidden states denoted as $\tilde{h}_i^n$. After applying the attention models, we further process each $\tilde{h}_i^n$ using a feed-forward network $FFN(.)$ defined in equation 3.7. This process yields a set of transformed hidden states on the final layer of encoder denoted as h_i^n and the decoder takes this representation for further process.

$$H_E^n = (FFN(\tilde{h}_1^n), \ldots, FFN(\tilde{h}_{n_x}^n)) \tag{3.3}$$

To be more specific, the $FFN(.)$ takes each $\tilde{h}_i^n$ as input and applies a non-linear transformation to it, resulting in a new set of hidden states denoted as H_E^n. The $FFN(.)$ function is applied independently to each element of $\tilde{h}_i^n$, producing corresponding elements in H_E^n. In other words, if $\tilde{h}_i^n$ has n_x elements, then H_E^n will also have n_x elements, where each element is obtained by applying the $FFN(.)$ function to the corresponding element in $\tilde{h}_{n_x}^n$. This process is repeated for each layer n in the encoder. At the final layer, the encoder outputs H_E^N, which represents the transformed hidden states from the last layer.

Step-3: Let $S_{<t}^n$ represent the hidden state of the n-th layer in the decoder prior to time step t. In other words, $S_{<t}^n$ consists of the hidden states (s_1^n, ... , s_{t-1}^n) up to the previous time

steps of t. i.e.,

$$S^n_{<t} = (s^n_1, \ldots, s^n_{t-1}) \tag{3.4}$$

It's important to note that S^0_1 is a special token that indicates the start of a sequence, and S^0_t represents the embedding of the predicted word at time step $t-1$.
At the n-th layer, we compute the updated hidden state $\hat{s}^n_t$ using the attention model $attn_S$. This attention model takes as input the previous hidden state s^{n-1}_t, as well as the sequences S^{n-1}_{t+1} and S^{n-1}_{t+1} and the whole process is computed as,

$$\begin{aligned}
\hat{s}^n_t &= attn_S(s^{n-1}_t, S^{n-1}_{<t+1}, S^{n-1}_{<t+1}); \\
\tilde{s}^n_t &= \frac{1}{2}(attn_{E,D}(\hat{s}^n_t, H^N_E, H^N_E) + attn_{LM,D}(\hat{s}^n_t, H_{LM}, H_{LM})), \\
s^n_t &= FFN(\tilde{s}^n_t)
\end{aligned} \tag{3.5}$$

The self-attention model, represented by $attn_S$, captures the relationships between different positions within a sequence. The BERT-decoder attention model, represented by $attn_{LM,D}$, focuses on the relevant information from the BERT output to aid in decoding. The encoder-decoder attention model, represented by $attn_{E,D}$, attends to the encoder representations and helps the decoder generate accurate translations the same as attention mechanism used in Vaswani et al. [46].
Equation 3.5 describes the iterative process of applying these attention models across multiple layers. By iterating through the layers, we obtain the final representation s^N_t, which captures the contextual information of the output sequence. This representation is then transformed using a linear function and passed through a softmax operation to obtain the predicted word $\hat{y}_t$ at time step t. The decoding process continues until an end-of-sentence token is generated.

In our framework, we utilize the output of BERT as an external representation of the input sequence. This allows us to leverage the pre-trained BERT model, regardless of the specific tokenization method used. To incorporate the BERT output into the neural machine translation (NMT) model, we employ an attention mechanism used by Zhu et al. [26]. This attention mechanism enables the NMT model to focus on the relevant information from BERT while generating translations, enhancing the overall translation quality.

Let $attn(q, K, V)$ represent the attention layer, where q, K, and V denote the query, key, and value, respectively. The query, q, is a vector of dimension d_q (where $d_q \in \mathbb{Z}$). The key set, K, and value set, V, both have the same cardinality, denoted by $|K| = |V|$. Each element, $k_i \in K$ and $v_i \in V$, is a vector of dimensions d_k/d_v, respectively. It is worth noting that d_q, d_k, and d_v

can have different values. The attention model can be described as follows:

$$attn(q, K, V) = \sum_{i=1}^{|V|} \alpha_i W_v v_i, \alpha_i = \frac{exp((W_q q)^T (W_k k_i))}{\sum_{i=1}^{|K|} exp((W_q q)^T (W_k k_i))} \tag{3.6}$$

where W_q, W_k and W_v are learnable parameters. We define the non-linear transformation we defined by Vaswani et al. [46] as follows:

$$FFN(x) = max(W_1 x + b_1, 0) W_2 + b_2 \tag{3.7}$$

In the given equation, the variable x represents the input, while W_1, W_2, b_1, and b_2 denote the parameters that need to be learned. The operator max is applied element-wise. Additionally, as introduced by Vaswani et al. [46] in the Transformer model, we also incorporate layer normalization.

While training, We use the drop-net technique introduced by Zhu et al. [26] to enhance the training process of our model. The primary motivation behind this technique is regularization, which helps to prevent overfitting and improve the generalization ability of the network. Our objective is to ensure that the features generated by BERT and the conventional encoder are fully utilized. The drop-net technique will have an impact on equations 3.2 and 3.5. Let us denote the drop-net rate as p_{net}, where p_{net} ranges from 0 to 1. During each training iteration, for any given layer n, we randomly sample a variable U^n from the interval $[0, 1]$. Then, the calculations for all $\tilde{h}_i^n$ in equation 3.2 are performed as follows:

$$\begin{aligned}
\tilde{h}_{i,dn}^n =& \mathbb{I}(U^n < \frac{p_{net}}{2}) \cdot attn_S(h_i^{n-1}, H_E^{n-1}, H_E^{n-1}) \\
&+ \mathbb{I}(U^n > 1 - \frac{p_{net}}{2}) \cdot attn_{LM,E}(h_i^{n-1}, H_{LM}, H_{LM}) \\
&+ \frac{1}{2}\mathbb{I}(\frac{p_{net}}{2} \leq U^n \leq 1 - \frac{p_{net}}{2}) \cdot (attn_S(h_i^{n-1}, H_E^{n-1}, H_E^{n-1}) \\
&+ attn_{LM,E}(h_i^{n-1}, H_{LM}, H_{LM}))
\end{aligned} \tag{3.8}$$

The $\mathbb{I}(\cdot)$ utilized in the equation is the indicator function. In each layer n, there is a probability of $\frac{p_{net}}{2}$ that either the BERT-encoder attention or self-attention is exclusively employed. With a probability of $(1 - p_{net})$, both attention models are utilized simultaneously. For instance, during a specific iteration, the first layer might exclusively use $attn_S$, while the second layer solely employs $attn_{LM,E}$. During inference time, we utilize the expected output of each attention model, which is calculated as the expectation of equation 3.2.

Similarly, we use the following drop-net trick for decoder when training:

$$\begin{aligned}\tilde{s}^n_{t,dn} =& \mathbb{I}(U^n < \frac{p_{net}}{2}) \cdot attn_{LM,D}(\hat{s}^n_t, H_{LM}, H_{LM}) \\ &+ \mathbb{I}(U^n > 1 - \frac{p_{net}}{2}) \cdot attn_{E,D}(\hat{s}^n_t, H^N_E, H^N_E) \\ &+ \frac{1}{2}\mathbb{I}(\frac{p_{net}}{2} \leq U^n \leq 1 - \frac{p_{net}}{2}) \cdot (attn_{LM,D}(\hat{s}^n_t, H_{LM}, H_{LM}) \\ &+ attn_{E,D}(\hat{s}^n_t, H^N_E, H^N_E))\end{aligned} \tag{3.9}$$

Again while inference time we use the same way as equation 3.5 for decoder.

3.4.3.2 Knowledge Distillation

This is where we tried to improve the performance of the BERT-Fused model in MT. As GPT is proven to outperform encoder only models like BERT in the task of text generation, we utilize GPT-2. During training, the Transformer model utilizes GPT as a reference or guide, predicting the next word based on the input context. This teacher-student approach helps the decoder of the Transformer model to generate more accurate and contextually appropriate word predictions.

A Transformer model is trained to learn parameters θ that can estimate the conditional likelihood $P_\theta(Y|X)$. This likelihood represents the probability of generating the target sequence Y given the input sequence X. The model is typically trained using Maximum Likelihood Estimation (MLE) or, equivalently, by minimizing the cross-entropy loss. MLE aims to find the parameters θ that maximize the likelihood of observing the actual target sequences in the training data, given the input sequences. Figure 9 shows the whole proposed model in one.

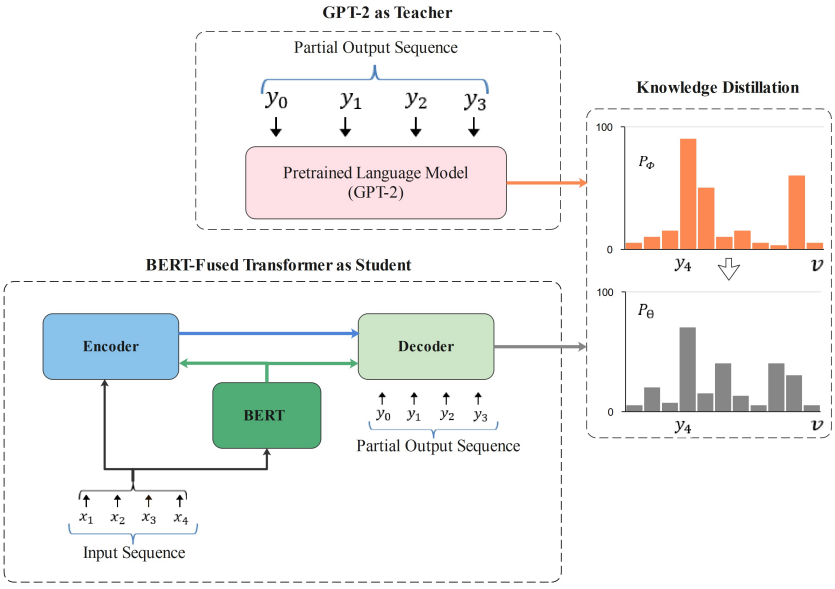


Figure 9. *BERT-GPT Incorporated Transformer model.*

During training, the model generates translations based on the input sequences and compares them to the ground truth translations. The cross-entropy loss measures the dissimilarity between the predicted translations and the actual translations. By minimizing this loss, the model adjusts its parameters θ to improve the quality of the generated translations, bringing them closer to the desired output. The equation is expanded as follows:

$$\begin{aligned} \mathcal{L}_{xe}(\theta) &= -logP_\theta(y|x) \\ &= -\sum_{t=1}^{N} logP_\theta(y_t|y_{<t}, x) \end{aligned} \tag{3.10}$$

Here, $P_\theta(y|x)$ represents the conditional probability of generating the translations y given the input sentences x, parameterized by θ. Taking the logarithm of this probability and negating it gives us the negative log-likelihood, which is commonly used as a loss function. The next step involves decomposing the conditional probability by considering the individual target words y_t in the translation sequence y. This is done using the chain rule of probability. In this summation, we calculate the negative log-likelihood for each target word in the translation sequence y, based on the previous words $y_{<t}$ and the input sentences x. The parameters θ are used to model the conditional distribution $P_\theta(y_t|y_{<t}, x)$.

The overall loss $\mathcal{L}_{xe}(\theta)$ is obtained by summing up the negative log-likelihoods for all target words in the translation sequence. By minimizing this loss function with respect to the parameters θ, machine translation models can be trained to generate more accurate translations that align closely with the ground truth translations.

In the context of knowledge distillation, the GPT-2 model can be seen as a teacher, while the Transformer model plays the role of a student. More specifically, the Transformer model can be trained using the following objective function:

$$\mathcal{L}_{GPT}(\theta) = -\sum_{w \in V} [P_\phi(y_t = w|y_{<t}) \cdot logP_\theta(y_t = w|y_{<t}, x)], \tag{3.11}$$

where the soft target probability $P_\phi(y_t)$ is estimated by the GPT-2 model with learned parameters ϕ, and V represents the output vocabulary. It's important to note that ϕ remains fixed throughout the distillation process. Figure 9 provides a visual representation of this learning process, where the goal is to align the word probability distribution $P_\theta(y_t)$ generated by the student with the distribution $P_\phi(y_t)$ provided by the teacher (referred to as distillation).

In order to enhance the performance of the Transformer student model, hard-assigned labels are also incorporated. The training of the final model involves utilizing a compound objective as

follows:

$$\mathcal{L}(\theta) = \alpha\mathcal{L}_{GPT}(\theta) + (1-\alpha)\mathcal{L}_{xe}(\theta), \tag{3.12}$$

where α is a hyperparameter used to adjust the relative significance of the two training targets: the soft-estimation derived from GPT model and the ground-truth hard-label. It's worth noting that our proposed approach imposes minimal constraints on the architecture of the integrated Transformer model.

By combining these two concepts, incorporating BERT for context-aware embedding and using GPT as a language generation and guiding model, we aim to improve the performance of the original Transformer model in MT.

3.4.4 Evaluation

The evaluation is a critical component of the research as it provides a means to measure the effectiveness and accuracy of the developed system. In this study, we use the BLEU (Bilingual Evaluation Understudy) score [70], chrF (Character F-score)[86], and TER (Translation Edit Rate)[72] as the evaluation metric for machine translation. BLEU is a widely used metric that measures the similarity between the machine-generated translations and the human-generated translations, with higher values indicating better translations. BLEU score is mathematically expressed as follows:

$$BLEU = min(\frac{output-len}{reference-len}, 1)(\prod_{i=1}^{4} precision_i)^{\frac{1}{4}} \tag{3.13}$$

chrF is a metric used to evaluate the quality of machine translation output. It measures the similarity between the reference translation and the machine-generated translation at the character level. The chrF metric is an extension of the BLEU metric, which operates at the n-gram level. While BLEU considers the precision of n-grams (contiguous sequences of n words) in the translations, chrF focuses on the precision and recall of character n-grams (sequences of n characters) in the translations. The chrF metric computes an F-score by combining the precision and recall values. The precision represents the ratio of correctly predicted character n-grams in the machine translation to the total number of character n-grams in the reference translation. The recall represents the ratio of correctly predicted character n-grams in the machine translation to the total number of character n-grams in the reference translation.

TER is a metric used for evaluating the quality of machine translation output. It operates at the sentence level and measures the number of edits required to transform the machine-generated

translation into the reference translation, which is considered the human-generated gold standard. TER considers four types of edits: substitutions, insertions, deletions, and shifts. Substitutions involve replacing one word or phrase with another, insertions involve adding an extra word or phrase, deletions involve removing a word or phrase, and shifts involve moving a word or phrase to a different position in the sentence. To calculate the TER score, the total number of edits required to align the machine translation with the reference translation is divided by the total number of words in the reference translation. This provides an indication of the level of editing needed to make the machine translation match the reference sentences.

To further evaluate the effectiveness of the proposed approach, we compare the performance of our approach with the state-of-the-art machine translation models which are the original Transformer model [46] and the approach used by Zhu et al. [26].

Chapter Four

4. Result Analysis

The Result and Analysis section aims to present and discuss the findings of our study on English-Afaan Oromo machine translation. In this section, we provide a detailed analysis of the experimental results, including the performance of the baseline Transformer model and the proposed model that incorporates monolingual pre-training techniques. We examine the translation quality, fluency, and semantic similarity achieved by the models in both the English-to-Afaan Oromo and Afaan Oromo-to-English translation directions. Furthermore, we discuss the implications of these results and analyze the strengths and weaknesses of the proposed model. Through this analysis, we gain insights into the effectiveness of pre-training techniques for improving machine translation in under-resourced languages and identify potential areas for future research and development.

4.1 Experimental Settings

For the dataset, we collected a parallel corpus consisting of English-Afaan Oromo sentence pairs for training and evaluation purposes. The dataset was carefully curated to ensure high-quality translations and to cover a diverse range of topics and sentence structures as far as we can. The dataset is collected from different sources including both primary and secondary sources and saved in the form of parallel sentences separated by tab. After pre-processing, we get 84,565 ready parallel sentences for training our model. We divided our available parallel dataset randomly into an 80/20 split for training purposes.

Table 7. *Train Test split for Parallel dataset.*

Type	Percentage	Sentences
Training	80%	67,652
Validation	10%	8,457
Test	10%	8,456
	Total	84,565

Table 7 shows the dataset used for Training (80%), validation (10%), and testing (10%) purposes for our model.

The machine translation models used in our experiments include a baseline Transformer model presented by Vaswani et al. [46] and the BERT-Fused model proposed by Zhu et al. [26] and our proposed model that incorporate BERT and GPT into original transformer model. The baseline

Transformer model served as a benchmark for comparison, while the proposed model aimed to leverage the benefits of pre-training techniques for improved translation quality.

In our study, we incorporate two essential concepts: language understanding and language generation. To address the language understanding aspect, we employ a transfer learning approach utilizing the $BERT_{BASE}$ model. We selected $BERT_{BASE}$ due to its strong performance and widespread usage in various natural language processing tasks. This model is a transformer-based language model that has been pre-trained on a large corpus of text data. It comprises a stack of 12 encoder layers, each with 12 self-attention heads, allowing it to capture contextual information effectively. The $BERT_{BASE}$ model is well-suited for knowledge understanding tasks as it can process and comprehend text input. In our experimental setup, we utilized the $BERT_{BASE}$ model's last layer and employed a masked language modeling (MLM) task to leverage its knowledge understanding capabilities. The MLM task involves predicting the masked tokens in the input sequence, enabling the model to learn contextual representations and effectively understand the knowledge encoded in the text. For fine-tuning, the Wordpiece tokenizer employed in our study is BertTokenizer [1]. The model we utilized is TFBertForMaskedLM [2], which is a BERT model equipped with a masked language modeling (MLM) head. This specific BERT model can only accept TensorFlow tensors. The checkpoint used for both the tokenizer and the model is bert-base-uncased.

```
model.summary()
```

```
Model: "tf_bert_for_masked_lm"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 bert (TFBertMainLayer)      multiple                  108891648

 mlm___cls (TFBertMLMHead)   multiple                  24459834

=================================================================
Total params: 109514298 (417.76 MB)
Trainable params: 109514298 (417.76 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 10. *Summary of MLM BERT model.*

Figure 10 examines the summary of the model. From the summary, we can see that the architecture consists of two distinct blocks of layers. The first block comprises the core BERT layer, which serves as the foundation for the model. The second block of layers is the MLM

[1]Available at https://huggingface.co/docs/transformers/model_doc/bert#transformers.BertTokenizer

[2]Available at https://huggingface.co/docs/transformers/v4.36.1/en/model_doc/bert#transformers.TFBertForMaskedLM

(Masked Language Model) head layer. The MLM model head layer utilizes the contextualized representations generated by the core BERT layer to make accurate predictions about the masked words, thereby enhancing the model's language understanding capabilities.

Building upon the solid foundation established by BERT for language understanding, we leverage the prowess of GPT-2 [15] into the task of language generation. Specifically, we fine-tune this pre-trained decoder-only transformer model for next word prediction, complementing BERT's knowledge extraction with GPT-2's fluency and target prediction ability.

Instead of masked language modeling, we employ a causal language modeling objective, where GPT-2 learns to predict the next word in a sequence. This approach optimizes the model for generating grammatically correct and continuations of the text. We prepared the dataset used for finetuning the dataset by tokenizing the Afan Oromo sentences from parallel dataset used for machine translate. This is because the vocabulary of GPT model and Transformer model should be the same in our context. We used the GPT2Tokenizer[3], ensuring compatible input for the model.

Fine-tuning involves carefully configuring the TFGPT2LMHeadModel[4] with specific optimizers and loss functions, followed by iterative training on monolingual data. Evaluating performance through perplexity and qualitative analysis of generated text allows us to assess the model's capability in creating meaningful and coherent continuations.

```
model.summary()
```

```
Model: "tfgpt2lm_head_model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 transformer (TFGPT2MainLay  multiple                  124439808
 er)

=================================================================
Total params: 124439808 (474.70 MB)
Trainable params: 124439808 (474.70 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 11. *Summary of GPT-2 model for NWP.*

Figure 11 provides a screenshoot of the summary of the GPT-2 model architecture used in our study. We employed the 124M trainable parameters model for next word prediction task in

[3]available at https://huggingface.co/docs/transformers/model_doc/gpt2#transformers.GPT2Tokenizer

[4]available at https://huggingface.co/docs/transformers/v4.36.1/en/model_doc/gpt2#transformers.TFGPT2LMHeadModel

auto-regressive fashion.

For the machine translation task, the architecture and setups we used are the same with settings presented by Vaswani et al. [46]. Implementing the Transformer model architecture, we developed an encoder-decoder framework. The architecture consisted of numerous layers, incorporating sub-layers like multi-head self-attention and feed-forward neural networks. The figure 12 below shows the summary of our proposed model.

```
transformer.summary()

Model: "transformer_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 encoder_3 (Encoder)         multiple                  144105216

 decoder_3 (Decoder)         multiple                  263890944

 dense_93 (Dense)            multiple                  18272978

=================================================================
Total params: 426269138 (1.59 GB)
Trainable params: 426269138 (1.59 GB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 12. *Summary of Proposed model.*

We made modifications to the model's structure to accommodate the dataset's scale and intricacy. Specifically, we employed six layers, 768-dimensional embeddings, and a 2048-dimensional internal dimension for the FeedForward layer. The number of self-attention heads remained unchanged at eight, as initially proposed in the original Transformer model Vaswani et al. [46]. We used the residual dropout Srivastava et al. [87] with dropout rate of 0.1 and we set the batch size to 64 to make sure fair comparison with our baseline models. Using a sequence-to-sequence (Seq2Seq) framework, we trained the model to minimize cross-entropy loss by comparing predicted target tokens with the actual target tokens in the dataset following the KD mechanism described in our paper under Chapter Three (Section 3.4.3.2). To optimize our model, we utilized the Adam optimizer [88] with the values of β_1, β_2 and ϵ set to 0.9, 0.98 and 10^{-9}. We implemented a custom learning rate scheduler based on the formula outlined in the original Transformer model [46].

$$lrate = d_{model}^{-0.5} * min(num\_step^{-0.5}, num\_step \cdot warmup\_steps^{-1.5}) \tag{4.1}$$

We used $warmup\_steps = 4000$ as it is. The figure 13 below shows the learning rate used for first $40,000$ steps while training. During training, we used a separate validation set to monitor performance and prevent overfitting. Finally, for inference, we utilized the trained

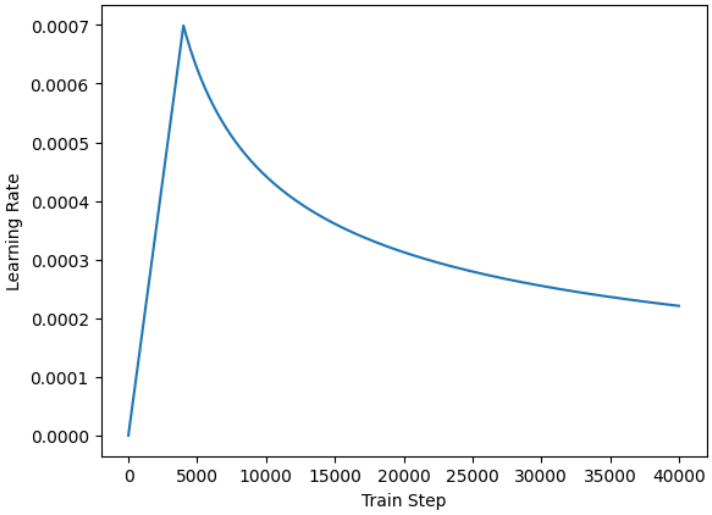


Figure 13. *Learning Rate for first 40,000 training steps.*

Transformer model by encoding source sentences, feeding them to the decoder to generate target translations autoregressively without any dependance of incorporated unsupervised learning language models by setting the p_{net} and α values to 0, so that the model translate the source sentence to target sentence independently during inference time.

To evaluate the performance of the models, we used standard evaluation metrics such as BLEU, chrF and TER. BLEU measures the n-gram overlap between the generated translations and the reference translations, while chrF is used to evaluate the quality of machine translation by comparing character n-grams between the reference and translated sentences. TER is a metric that quantifies the number of edits (insertions, deletions, substitutions) required to transform the translated sentence into the reference sentence. It provides a measure of the overall accuracy and fluency of the translation. Lower TER scores indicate better translation quality. We utilize a more sophisticated implementation of these evaluations from detokenized sacreBLEU[5] for fair comparison with previous works. We use These metrics provide quantitative measures to assess translation accuracy, fluency, and semantic similarity of translated sentences.

The experimentation process involved utilizing Tensorflow-2.15.0 as the core framework for our source code. To ensure accessibility and flexibility, we conducted our training on both the resourceful Google Colab Pro and the free versions of the platform.

4.2 Experimental Results

In our experiment, we compared two baseline models with our proposed model. To ensure a fair comparison, we placed all the models in the same environment. The table 8 below shows the

[5]available at `https://github.com/mjpost/sacreBLEU`

number of trainable parameters for each model:

Table 8. *Experimented models trainable parameters and training time taken.*

Models	Trainable Parameters	Training Time
Transformer	307M	4 Hr
BERT-Fused	426M	9 Hr
Proposed	426M	12 Hr

From the table, we can see that the Transformer model has 307 million trainable parameters, while both the BERT-Fused model and our proposed model have 426 million trainable parameters. The reason for the increase in parameters in our proposed model is because we introduced an additional attention mechanism to incorporate the output of BERT model into the transformer model for context understanding purpose.

There is no difference in the number of parameters between the BERT-Fused model and our proposed model. However, there is a difference in training time, with our proposed model taking approximately three hours longer. This is because our model uses a knowledge distillation mechanism, which is time-consuming for predicting the next word in the Teacher force approach introduced in our paper.

4.2.1 English-to-Afaan Oromo Translation

The results indicate a significant improvement in translation quality when using the proposed model. The BLEU score increased from 35.75 for the baseline model to 42.09for the proposed model, indicating a notable enhancement in translation accuracy, fluency, and semantic similarity with the reference sentences. The following table 9 summerizes the evaluation results for proposed model with baselines (Transformer [46] and BERT-Fused [26]).

Table 9. *English-to-Afaan Oromoo translation evaluation results.*

Models	BLEU Score (%)	chrF Score	TER Score
Transformer	32.82	65.08	35.5
BERT-Fused	35.75	68.56	29.06
Proposed	42.09	73.97	21.64

The table presents the results of English to Afaan Oromo translation for three models: the original Transformer model, the BERT-Fused model, and the proposed model.

Starting with the original Transformer model, it achieved a baseline BLEU score of 32.82, indicating moderate translation quality. However, there is room for improvement in accurately

capturing the meaning and structure of the reference translations. The chrF score of 65.08 suggests a moderate level of character-level similarity with the reference translations, while the TER score of 35.5 implies a relatively high error rate.

Moving on to the BERT-Fused model, it showed improved performance compared to the original Transformer model. The BLEU score increased to 35.75, indicating enhanced translation quality with better n-gram overlap. The chrF score also increased to 68.56, indicating a higher character-level similarity. Moreover, the TER score decreased to 29.06, suggesting a reduced error rate and improved translation accuracy.

The proposed model, which leveraged unsupervised pre-training techniques, outperformed both baselines. It achieved a significantly higher BLEU score of 42.09, indicating substantial improvement in translation quality compared to the original Transformer model. The chrF score further increased to 73.97, demonstrating a higher character-level similarity. Additionally, the TER score decreased to 21.64, indicating a substantial decrease in error rate and improved translation accuracy.

Specific examples of translations generated by the models illustrate the improvements achieved. The baseline model often struggled with complex sentence structures and idiomatic expressions, resulting in inaccuracies and loss of meaning. However, the proposed model demonstrated a better understanding of the source text, leading to more accurate and contextually appropriate translations.

4.2.2 Afaan Oromo-to-English Translation

Similar to the English-to-Afaan Oromo translation, the proposed model showed significant improvements in translation quality. The BLEU score increased from 40.35 for the baseline model to 44.51 for the BERT-incorporated model, indicating enhanced translation accuracy. The

Table 10. *Afaan Oromoo-to-English translation evaluation results.*

Models	**BLEU Score (%)**	**chrF Score**	**TER Score**
Transformer	34.15	68.57	30.42
BERT-Fused	40.35	75.43	25.09
Proposed	44.51	79.66	18.91

table 10 presents the results of Afaan Oromo to English translation for three models: the original Transformer model, the BERT-Fused model, and the proposed model leveraging unsupervised pre-training techniques.

Starting with the original Transformer model, it achieved a baseline BLEU score of 34.15, indicating moderate translation quality. However, similar to the English to Afaan Oromo translation, there is room for improvement in accurately capturing the meaning and structure of the reference translations. The chrF score of 68.57 suggests a moderate level of character-level similarity with the reference translations, while the TER score of 30.42 implies a relatively high error rate.

When we come to the BERT-Fused model, it showed improved performance compared to the original Transformer model. The BLEU score boosted to 40.35, indicating enhanced translation quality with better n-gram overlap. The chrF score also increased to 75.43, indicating a higher character-level similarity. Moreover, the TER score decreased to 25.09, suggesting a reduced error rate and improved translation accuracy.

The proposed model, leveraging unsupervised pre-training techniques, outperformed also English-to-Afaan Oromo translation in both baselines. It achieved a significantly higher BLEU score of 44.51, indicating substantial improvement in translation quality compared to the original Transformer model and BERT-Fused model. The chrF score further increased to 79.66, demonstrating a higher character-level similarity. Additionally, the TER score decreased to 18.91, indicating a substantial decrease in error rate and improved translation accuracy.

Examining specific translation examples, we observed that the baseline model struggled with capturing the nuances of Afaan Oromo sentence structures and idiosyncrasies, often resulting in grammatical errors and incorrect word choices. However, the proposed model demonstrated a better grasp of the Afaan Oromo language, producing more precise and contextually appropriate translations.

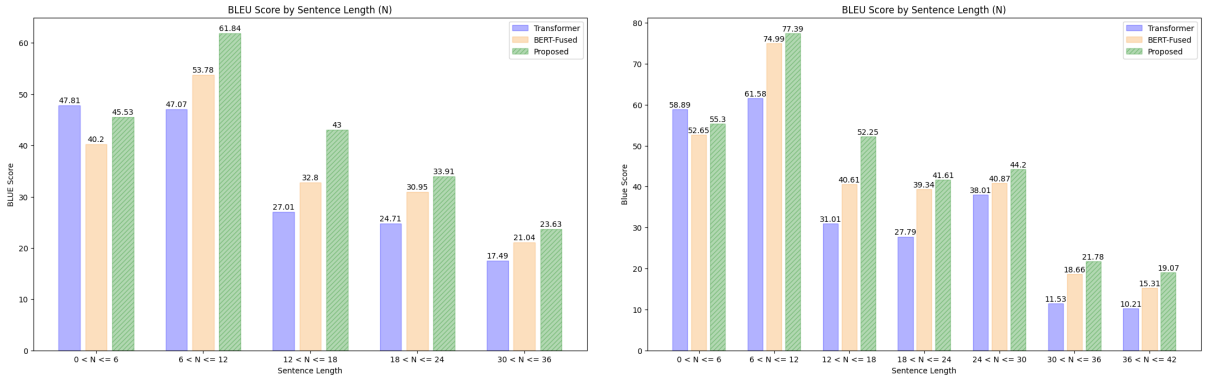
4.3 Comparative Analysis

Under this section we will discuss the comparative analysis for translation directions (Afaan Oromo-to-English and English-to-Afaan Oromo translations), the use of different hyper-parameters, and the use of monolingual data.

As we have discussed in previous section in table 9 and 10 the results indicate that both translation directions benefited from the proposed model, but there were slight variations in performance. The English-to-Afaan Oromo translation exhibited slightly higher improvement rates in terms of BLEU and TER scores compared to the Afaan Oromo-to-English translation. This suggests that the pre-training models, particularly GPT-2, provided better support for generating and translating from English to Afaan Oromo. The superior performance in generating translations from English to Afaan Oromo can be attributed to the fact that the BERT model is pretrained on

an extensive corpus of English text. This extensive pre-training allows the model to develop a deeper understanding of English sentences, enabling it to generate higher-quality contextual representations from its language model. Moreover, the inclusion of GPT-2 as a supporting model created further variations of the translation process. GPT-2 is known for its ability to generate coherent and contextually appropriate text. By leveraging GPT-2 during translation, the student model benefits from GPT-2's proficiency in generating high-quality words and phrases for Afaan Oromo, resulting in improved translation quality from English to Afaan Oromo.

The comparative analysis also highlights the impact of pre-training techniques on translation quality and fluency. The incorporation of BERT and GPT in the proposed model improved the handling of complex sentence structures, idiomatic expressions, and linguistic nuances in both translation directions. This demonstrates the effectiveness of leveraging pre-training techniques for addressing language-specific challenges and improving the overall translation performance. The figure 14 below shows the effect of the proposed model on sentence length. Upon analyzing both Figure 14a and Figure 14b, a significant observation emerges. The BERT-



(a) (b)

Figure 14. *BLEU Scores of different output sentence lengths 14a for Afaan Oromo translated sentences and 14b for English translated sentences.*

Fused model consistently encounters difficulties in generating short sentences, particularly sentences with less than four words. This can be attributed to the contextual understanding provided by the BERT model, which leads the transformer model to excessively interpret the words and introduces unnecessary gaps that prompt the model to consider irrelevant terms during translation. In contrast, the original transformer model surpasses both our proposed model and the BERT-Fused model without any supplementary information. Regarding our proposed model, it exhibits slightly lower performance than the original transformer model. However, it still outperforms the BERT-Fused model by a substantial margin. This suggests that incorporating knowledge distillation (KD) into the language generation process aids the model in selecting more appropriate words during translation, thus improving the overall quality of the generated text.

Moreover, in contrast to the baseline models, our proposed model exhibited superior performance in all translation directions, particularly when dealing with longer sentences. This is evident from the figures, which clearly demonstrate that the original transformer model struggles when confronted with larger sentences, specifically in the translation of sentences containing more than twelve words in both directions. These observations indicate that our proposed model leveraged the advantages of pre-training models to effectively handle longer sentences. However, upon examining the figure in its entirety, it becomes evident that there is still homework for improvement in effectively handling longer sentences.

In the context of hyper-parameter analysis, many of the hyper-parameters are typically determined based on recommendations from previous studies. Additionally, certain initial studies were conducted and incorporated into this study, although they are not specifically discussed in this section. In our research, we focused on investigating two specific hyper-parameters: the P_{net} value for BERT-fusion and the α value for knowledge distillation. Interestingly, our findings align with the results suggested by Zhu et al. [26] and Chen et al. [28]. Specifically, we determined that setting the P_{net} value to 1.0 and the α value to 0.1 yielded the same outcome as indicated in their respective papers.

Focusing on our proposed model, we conducted additional experiments to investigate the impact of different datasets. Specifically, we trained the model using two different approaches: (1) utilizing our collected Afaan Oromo monolingual data described in Section 3.2, and (2) exclusively using sentences from the parallel dataset. We performed these experiments separately for both the Afaan Oromo-to-English and English-to-Afaan Oromo translation directions.

For Afaan Oromo-to-English translation, we fine-tuned the BERT model using Afaan Oromo monolingual data, as well as Afaan Oromo sentences from the parallel dataset. Additionally, we fine-tuned the GPT-2 model for next-word prediction using English sentences from the parallel dataset. Similarly, for English-to-Afaan Oromo translation, we fine-tuned the BERT model using English sentences from the parallel dataset, while utilizing the pretrained BERT model directly for English sentences. As for GPT-2, we fine-tuned it using both sentences exclusively from the parallel dataset and the monolingual Afaan Oromo corpus, ensuring the same vocabulary size with the transformer model. These different configurations were individually incorporated into the training of the transformer model, allowing us to assess their respective impacts. The tables 11 and 12 below summarize the results on all experiments on different datasets.

From the analysis of table 11, it is evident that incorporating the knowledge acquired by BERT through fine-tuning it on Afaan Oromo sentences from the parallel dataset, along with the next-word prediction (NWP) learned by GPT-2 on English sentences from the same dataset, using the same vocabulary as the transformer model, has yielded a translation BLEU score of 39.7. This

Table 11. *PTLMs fine-tuned for Afaan Oromoo-to-English translation. (SPD: Sentences from Parallel Dataset, MoLD: Monolingual Dataset)*

PTLMs incorporated to transformer model	BLEU Score (%)
BERT + SPD and GPT-2 + SPD	39.7
BERT + MoLD and GPT-2 + SPD	**44.51**

approach, where the GPT-2 model served as the teacher for knowledge distillation (KD) purposes for transformer model (the student), outperformed the baseline original transformer model. However, it fell short of the performance achieved by the BERT-Fused model, which was trained on Afaan Oromo monolingual data and incorporated into the transformer model for context understanding, yielding a BLEU score of 40.35. This suggests that the sentences available exclusively in the parallel dataset are insufficient for comprehensive context understanding.

In the second experiment, fine-tuning the BERT model on Afaan Oromo monolingual data, while utilizing the GPT-2 model fine-tuned on English sentences from the parallel dataset, resulted in a notable improvement. The BLEU score increased to 44.51, surpassing both the transformer and BERT-Fused models. This experiment highlights the significant impact of incorporating monolingual data in enhancing the context understanding task. Furthermore, comparing the BERT-Fused model and our proposed model, both trained on the same fine-tuned BERT model using monolingual data, reveals a 4.16 increase in BLEU score for our proposed model. This demonstrates how the teacher model (GPT-2) aided the transformer model in generating accurate words and phrases during text generation.

Table 12. *PTLMs fine-tuned for English-to-Afaan Oromoo translation. (SPD: Sentences from Parallel Dataset, MoLD: Monolingual Dataset, the bracket tells vocabulary used by the student model.)*

PTLMs incorporated to transformer model	BLEU Score (%)
BERT and GPT-2 + SPD	38.02
BERT and GPT-2 + MoLD (Vocabulary from SPD)	30.46
BERT and GPT-2 + MoLD (Vocabulary from MoLD)	33.51
BERT + SPD and GPT-2 + SPD	41.58
BERT + SPD and GPT-2 + MoLD (Vocabulary from SPD)	32.03
BERT + SPD and GPT-2 + MoLD (Vocabulary from MoLD)	**42.09**

Analyzing the results presented in table 12, which pertains to English-to-Afaan Oromoo translation, we employed six different strategies to train the transformer model.

In the first strategy, we trained the model by directly incorporating the pre-trained BERT model for context representation of English sentences, while fine-tuning the GPT-2 model on Afaan

Oromoo sentences from the parallel dataset. This yielded a BLEU score of 38.02, surpassing both of our baseline models.

Moving on to the second strategy, the model was trained by directly incorporating the pre-trained BERT model, while the GPT-2 model was fine-tuned on a dataset consisting of monolingual data with the same vocabulary size as the student model (the student model trained by using vocabulary on sentences from parallel dataset and the teacher GPT model's vocabulary size limited to student model's vocabulary size). However, this resulted in a lower BLEU score of 30.46. The decline in performance can be attributed to the misleading nature of the teacher model, which was trained on sentences that were out of the vocabulary (OOV) of the transformer model. In our third experiment, focusing on second experiment's gap, we retrained our student model by creating the same vocabulary with teacher model fine-tuned on the monolingual data, the result has shown increase in performance. The BLEU score increased to 33.51surpassed the original transformer model but still underperformed when compared with our second baseline model which is BERT-Fused model. The reason BERT-Fused model outperformed our third experiment is that the BERT model used in our second baseline is the fine-tuned BERT model. From this result we can understand that the choice of vocabulary size has great impact on the model's performance. At the same time, fine-tuning BERT on available dataset rather than directly using the pre-trained one has shown great performance (even if BERT model is pre-trained on large English monolingual corpus).

For the fourth, fifth and sixth experiments, we fine-tuned the BERT model using English sentences from the parallel dataset. In the fourth experiment, we utilized the fine-tuned GPT-2 model on Afaan Oromo sentences from the parallel dataset, with the same vocabulary as the student model, resulting in an improved BLEU score of 41.58. However, in the fifth experiment, when we used the fine-tuned GPT-2 model on the monolingual dataset for next-word prediction, the BLEU score decreased to 32.03. This decrease in result happened because of the teacher GPT model introduces new vocabulary to the student model. In this experiment, the vocabulary for teacher (finetuned on MoLD) and student (trained on SPD) models are not the same, but vocabulary size is the same for both of them. In knowledge distillation concept the vocabulary size for teacher and student models has to be equal (in size) [28]. Considering the above limitation and decrease in performance, we conducted the sixth experiment. We used BERT fine-tuned on English sentences from parallel dataset for language understanding and we used GPT fine-tuned on next-word-prediction by using large monolingual Afaan Oromo dataset for language generation, we get BLEU score of 42.09. While training, we created Afaan Oromo vocabulary from MoLD for student transformer model. Surprisingly, the performance of the model outperformed all other our experiments. From this result we reached on the conclusion of using large monolingual data is advantageous and the use of the same vocabulary for student and teacher models leads to great performance.

From these experiments, we can draw two important conclusions. Firstly, the advantage of fine-tuning the BERT model on the available dataset rather than directly using the pre-trained one is evident. Secondly, using the same vocabulary for both the teacher and student models in the knowledge distillation (KD) task proves to be advantageous.

4.4 Discussion of the Results of the Study

The results of the study demonstrated significant improvements in translation quality by incorporating the unsupervised pre-training techniques. The integration of BERT's language understanding and GPT-2's language generation capabilities into the Transformer model led to more accurate and fluent translations in both directions. The system exhibited a better grasp of the source language context, resulting in enhanced comprehension and accurate representation of the input sentence. Moreover, the knowledge distillation approach using GPT-2 provided valuable guidance for the Transformer model to generate more contextually appropriate translations. Specially it helped the model in translating long sentences. Based on the findings in this study we discuss the research questions of this study as follows.

RQ1: How can we marry the concepts of unsupervised language understanding and language generation to get better translation quality?

To achieve better translation quality, the study emphasizes the importance of integrating the concepts of unsupervised language understanding and language generation. The pre-training techniques used in this work, BERT for language understanding and GPT-2 for language generation, play a crucial role in marrying these concepts.

By fine-tuning BERT on MLM using monolingual data, the model acquires a deep understanding of the source language. This understanding is then integrated into the Transformer model through an additional attention mechanism, enabling the system to better comprehend the context and nuances of the input sentence. On the other hand, GPT-2 is employed as a knowledge distillation approach, where the word predictions generated by GPT-2 act as soft labels. This allows the Transformer model to refine its language generation capabilities and produce more accurate translations. The study successfully marries unsupervised language understanding and language generation techniques through the integration of BERT and GPT-2 in the Transformer model. By looking into the experimental results, we can see that the language understanding and language generation models supported each other in translating high quality translation.

RQ2: How does the incorporation of pre-training techniques, such as BERT and GPT, in transformer model improve the quality of English-Afaan Oromo machine translation?

The findings of the study illustrate the effectiveness of monolingual data results in improved translation quality. By leveraging unsupervised pre-training techniques fine-tuned on monolingual data, the system gains a better grasp of the source language context, leading to enhanced comprehension and generation of more accurate and contextually appropriate translations. By leveraging large amounts of monolingual data, BERT was fine-tuned on the Masked Language Modeling (MLM) objective, enabling it to acquire a deep understanding of the source language. This understanding was then integrated into the Transformer model through an additional attention mechanism, resulting in improved translation quality. At the same time, GPT was fine-tuned on monolingual data for next word prediction (NWP) task and supported the transformer model in generating fluent and contextually appropriate translation of target language. The system demonstrated enhanced context comprehension, fluency, and the ability to handle linguistic nuances, leading to more accurate and natural translations in both translation directions. These findings underscore the potential of unsupervised pre-training techniques for improving machine translation performance in low-resource language scenarios by using monolingual data.

The comparison between the incorporation of pre-training techniques in transformer model and the original transformer model for low-resource languages reveals significant advantages of the pre-training techniques. The original transformer model heavily rely on parallel corpora, which are often scarce or unavailable for low-resource languages like Afaan Oromo. From our experiment this transformer model always suffer in understanding and generating long sentences. This problem is not only from the model, but from lack of enough dataset. If this model is trained on large parallel corpora it is worth nothing that it can handle long sentences also. But the problem is there is no enough parallel dataset and it is very expensive to prepare it for under resourced languages (Afaan Oromo in this context). In contrast, the unsupervised pre-training techniques employed in this study leverage monolingual data, which is more readily available. By fine-tuning BERT on MLM and GPT-2 on next word prediction (NWP) within the same vocabulary size of the transformer model, the system effectively captures language understanding and generation capabilities.

However, it is important to note that despite the improvements achieved, there are still limitations to be addressed. The scarcity of parallel data remains a challenge for Afaan Oromo, which can impact the generalizability of the models. Additionally, the proposed model may still encounter difficulties in translating domain-specific or technical content accurately. Further research and development efforts should focus on augmenting the available parallel dataset through data synthesis techniques, or exploring transfer learning methods from related languages. Domain-specific adaptation and fine-tuning techniques should also be investigated to improve the models' performance in specialized domains. Moreover, conducting evaluations on real-world data, including user-generated content or specific domains, would provide a more comprehensive understanding of the system's performance and its practical applicability.

In conclusion, the experimental results demonstrate the effectiveness of incorporating pre-training techniques, specifically BERT and GPT, into Transformer model for English-Afaan Oromo machine translation. The proposed model significantly improve translation accuracy, fluency, and semantic similarity in both translation directions. This research contributes to the advancement of under-resourced language processing and provides insights for future studies aiming to bridge the language barrier for Afaan Oromo speakers.

Chapter Five

5. Conclusion and Recommendation

5.1 Conclusion

In conclusion, this study aimed to enhance English-Afaan Oromo machine translation through pre-training in Transformer models using unsupervised learning techniques. By incorporating state-of-the-art pre-trained models, BERT and GPT, into the original Transformer model, we successfully improved translation accuracy, fluency, and semantic similarity.

Through extensive experimentation and analysis, we observed significant performance improvements in both the English-to-Afaan Oromo and Afaan Oromo-to-English translation directions. The proposed model consistently outperformed the baseline model, showcasing the effectiveness of leveraging pre-training techniques in under-resourced language translation tasks.

The findings highlight the importance of unsupervised pre-training language models on large amounts of unlabeled data to capture the underlying linguistic patterns and improve translation quality. By leveraging the knowledge acquired during pre-training, the machine translation system demonstrated enhanced understanding of context, improved handling of linguistic nuances, and better fluency in generating translations.

Ultimately, this integration of BERT and GPT-2 demonstrates the synergy between language understanding and generation, paving the way for advanced NLP solutions in Afaan Oromo and other under-resourced languages. The future holds great promise for pushing the boundaries of what these models can achieve, and every work with direction plays a crucial role in this fascinating journey.

5.2 Recommendations

Based on the findings and limitations encountered during the research, the following recommendations are provided for future work:

Dataset Augmentation: The scarcity of parallel data remains a challenge for under-resourced languages like Afaan Oromo. Future research should focus on augmenting the available parallel dataset through data synthesis techniques, leveraging monolingual data, or exploring transfer learning methods from related languages.

Model Complexity (Time and Space): Although our primary focus was on enhancing transla-

tion quality, we acknowledge that model complexity in terms of time and space requirements is an important consideration. In future research, it is recommended to investigate and optimize the model's complexity to reduce its size and training time. Exploring techniques such as model compression, parameter sharing, or efficient architectures can help mitigate the additional resource demands of our proposed models, allowing for more practical deployment.

Domain-specific Adaptation: While the proposed model showed promising results, further investigation into domain-specific adaptation is recommended. Incorporating domain-specific parallel data or fine-tuning techniques may lead to improved performance in specialized domains such as technical or medical translations.

Evaluation on Real-world Data: Conducting evaluations on real-world data, including user-generated content or specific domains, would provide a more comprehensive understanding of the system's performance and its practical applicability in real-world scenarios.

Collaborative Efforts: Collaboration among researchers, language experts, and native speakers of Afaan Oromo is crucial for building robust and accurate machine translation systems. Collaborative efforts can help address language-specific challenges, refine translation models, and ensure the availability of high-quality language resources. Finally, by considering other NLP tasks that touches MT directly or indirectly such as word sense disambiguation [89] can help the MT task to get better result.

By pursuing these recommendations, we can further advance the field of English-Afaan Oromo machine translation, contribute to the development of under-resourced language processing, and ultimately bridge the language barrier for Afaan Oromo speakers.

References

[1] Daniel Jurafsky and James H Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.

[2] Mary Yoko Brannen, Rebecca Piekkari, and Susanne Tietze. "The multifaceted role of language in international business: Unpacking the forms, functions and features of a critical challenge to MNC theory and performance". In: *Language in international business: Developing a field* (2017), pp. 139–162.

[3] Emma Steigerwald et al. "Overcoming language barriers in academia: Machine translation tools and a vision for a multilingual future". In: *BioScience* 72.10 (2022), pp. 988–998.

[4] Ebrahim Ansari et al. "Findings of the IWSLT 2020 evaluation campaign". In: *Proceedings of the 17th International Conference on Spoken Language Translation*. 2020, pp. 1–34.

[5] Martin Kay. "The proper place of men and machines in language translation". In: *machine translation* 12 (1997), pp. 3–23.

[6] Thi-Tam Nguyen et al. "Rule-Based Machine Translation for the Automatic Translation of Vietnamese Sign Language". In: *International Journal of Language and Linguistics* 11.6 (2023), pp. 191–198.

[7] Yu Shiwen and Bai Xiaojing. "Rule-based machine translation". In: *Routledge encyclopedia of translation technology*. Routledge, 2014, pp. 186–200.

[8] Peter F Brown et al. "The mathematics of statistical machine translation: Parameter estimation". In: (1993).

[9] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[10] Mary Hearne and Andy Way. "Statistical machine translation: a guide for linguists and translators". In: *Language and Linguistics Compass* 5.5 (2011), pp. 205–226.

[11] Anup Kumar Barman et al. "Word Sense Disambiguation applied to Assamese-Hindi Bilingual Statistical Machine Translation". In: *Engineering, Technology & Applied Science Research* 14.1 (2024), pp. 12581–12586.

[12] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[14] Yonghui Wu et al. "Google's neural machine translation system: Bridging the gap between human and machine translation". In: *arXiv preprint arXiv:1609.08144* (2016).

[15] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[16] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[17] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[18] Kaiming He, Ross Girshick, and Piotr Dollár. “Rethinking imagenet pre-training”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 4918–4927.

[19] Qizhe Xie et al. “Self-training with noisy student improves imagenet classification”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10687–10698.

[20] Sameer Bansal et al. “Pre-training on high-resource speech recognition improves low-resource speech-to-text translation”. In: *arXiv preprint arXiv:1809.01431* (2018).

[21] Yung-Sung Chuang et al. “Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering”. In: *arXiv preprint arXiv:1910.11559* (2019).

[22] Daniel S Park et al. “Specaugment: A simple data augmentation method for automatic speech recognition”. In: *arXiv preprint arXiv:1904.08779* (2019).

[23] Surangika Ranathunga et al. “Neural machine translation for low-resource languages: A survey”. In: *ACM Computing Surveys* 55.11 (2023), pp. 1–37.

[24] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. “A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings”. In: *arXiv preprint arXiv:1805.06297* (2018).

[25] Guillaume Lample et al. “Unsupervised machine translation using monolingual corpora only”. In: *arXiv preprint arXiv:1711.00043* (2017).

[26] Jinhua Zhu et al. “Incorporating bert into neural machine translation”. In: *arXiv preprint arXiv:2002.06823* (2020).

[27] Matthew E Peters et al. “Knowledge enhanced contextual word representations”. In: *arXiv preprint arXiv:1909.04164* (2019).

[28] Yen-Chun Chen et al. “Distilling knowledge learned in BERT for text generation”. In: *arXiv preprint arXiv:1911.03829* (2019).

[29] Qihao Zhu and Jianxi Luo. “Generative pre-trained transformer for design concept generation: an exploration”. In: *Proceedings of the Design Society* 2 (2022), pp. 1825–1834.

[30] Sehar Khan. *10 Most Spoken Languages in Africa*. 2018. URL: https://www.marstranslation.com/blog/10-most-spoken-languages-in-africa (visited on 04/25/2023).

[31] Ethnologue. *Oromo*. 2021. URL: https://www.ethnologue.com/language/orm/ (visited on 04/25/2023).

[32] Wondimu Tegegne. "The development of written Afan Oromo and the appropriateness of Qubee, Latin Script, for Afan Oromo writing". In: *Historical Research Letter* 28 (2016).

[33] Adere Gari Waldeyes. "Bi-Directional English to Afaan Oromo Neural Machine Translation using a Deep Learning Approach". PhD thesis. 2022.

[34] Debela Tesfaye. "A rule-based Afan Oromo Grammar Checker". In: *International Journal of Advanced Computer Science and Applications* 2.8 (2011).

[35] Debela Tesfaye. "Designing a stemmer for afaan oromo text: A hybrid approach". PhD thesis. Addis Ababa University, 2010.

[36] Teferi Degeneh Bijiga. *The development of Oromo writing system*. University of Kent (United Kingdom), 2015.

[37] Alan M Turing. "Can a machine think". In: *Mind* 59.236 (1950), pp. 433–460.

[38] Marta R Costa-jussà et al. "No language left behind: Scaling human-centered machine translation". In: *arXiv preprint arXiv:2207.04672* (2022).

[39] Dorothy Kenny. "Human and machine translation". In: *Machine translation for everyone: Empowering users in the age of artificial intelligence* 18 (2022), p. 23.

[40] Erik Brynjolfsson et al. "Does machine translation affect international trade? Evidence from a large digital platform". In: *Management Science* 64.11 (2018), pp. 4543–4560.

[41] Warren Weaver. *Typescript, July 1949. Translation. Reprinted in Locke, William N., and A. Donald Booth (eds) Machine translation of languages: fourteen essays*. 1955.

[42] John Hutchins. "Commercial systems". In: *Computers and Translation: A translator's guide* 35 (2003), p. 161.

[43] John Hutchins. "ALPAC: the (in) famous report". In: *Readings in machine translation* 14 (2003), pp. 131–135.

[44] Martin Kay. "The proper place of men and machines in language translation (Report CSL-80-11)". In: *Palo Alto, CA: Xerox Corporation* (1980).

[45] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27 (2014).

[46] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[47] Chris Callison-Burch et al. “Findings of the 2011 workshop on statistical machine translation”. In: *Proceedings of the sixth workshop on statistical machine translation*. 2011, pp. 22–64.

[48] Kyunghyun Cho et al. “On the properties of neural machine translation: Encoder-decoder approaches”. In: *arXiv preprint arXiv:1409.1259* (2014).

[49] Mia Xu Chen et al. “The best of both worlds: Combining recent advances in neural machine translation”. In: *arXiv preprint arXiv:1804.09849* (2018).

[50] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. “Effective approaches to attention-based neural machine translation”. In: *arXiv preprint arXiv:1508.04025* (2015).

[51] Jiajun Zhang and Chengqing Zong. “Neural machine translation: Challenges, progress and future”. In: *Science China Technological Sciences* 63.10 (2020), pp. 2028–2050.

[52] Paul J Werbos. “Generalization of backpropagation with application to a recurrent gas market model”. In: *Neural networks* 1.4 (1988), pp. 339–356.

[53] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[54] Jeff Heaton. “Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618”. In: *Genetic programming and evolvable machines* 19.1-2 (2018), pp. 305–307.

[55] Mikel Artetxe et al. “A call for more rigor in unsupervised cross-lingual learning”. In: *arXiv preprint arXiv:2004.14958* (2020).

[56] Yunsu Kim, Yingbo Gao, and Hermann Ney. “Effective cross-lingual transfer of neural machine translation models without shared vocabularies”. In: *arXiv preprint arXiv:1905.05475* (2019).

[57] Baijun Ji et al. “Cross-lingual pre-training based transfer for zero-shot neural machine translation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01. 2020, pp. 115–122.

[58] Yunsu Kim et al. “Pivot-based transfer learning for neural machine translation between non-English languages”. In: *arXiv preprint arXiv:1909.09524* (2019).

[59] Shuo Ren et al. “Explicit cross-lingual pre-training for unsupervised machine translation”. In: *arXiv preprint arXiv:1909.00180* (2019).

[60] Aizhan Imankulova et al. “Exploiting out-of-domain parallel data through multilingual transfer learning for low-resource neural machine translation”. In: *arXiv preprint arXiv:1907.03060* (2019).

[61] Tom Kocmi and Ondřej Bojar. “Trivial transfer learning for low-resource neural machine translation”. In: *arXiv preprint arXiv:1809.00357* (2018).

[62] Alexis Conneau et al. “Unsupervised cross-lingual representation learning for speech recognition”. In: *arXiv preprint arXiv:2006.13979* (2020).

[63] Surafel M Lakew et al. “Multilingual neural machine translation for low-resource languages”. In: *IJCoL. Italian Journal of Computational Linguistics* 4.4-1 (2018), pp. 11–25.

[64] Raj Dabre, Atsushi Fujita, and Chenhui Chu. “Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 1410–1416.

[65] Wenbo Zhang et al. “Pre-training on mixed data for low-resource neural machine translation”. In: *Information* 12.3 (2021), p. 133.

[66] Zihan Liu, Genta Indra Winata, and Pascale Fung. “Continual mixed-language pre-training for extremely low-resource neural machine translation”. In: *arXiv preprint arXiv:2105.03953* (2021).

[67] Colin Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 5485–5551.

[68] Mattia A Di Gangi and Marcello Federico. “Monolingual embeddings for low resourced neural machine translation”. In: *Proceedings of the 14th International Conference on Spoken Language Translation*. 2017, pp. 97–104.

[69] Atnafu Lambebo Tonja et al. “Low-Resource Neural Machine Translation Improvement Using Source-Side Monolingual Data”. In: *Applied Sciences* 13.2 (2023), p. 1201.

[70] Kishore Papineni et al. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[71] Satanjeev Banerjee and Alon Lavie. “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72.

[72] Matthew Snover et al. “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*. 2006, pp. 223–231.

[73] George Doddington. “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics”. In: *Proceedings of the second international conference on Human Language Technology Research*. 2002, pp. 138–145.

[74] Ondřej Bojar et al. “Findings of the 2017 conference on machine translation (wmt17)”. In: Association for Computational Linguistics. 2017.

[75] Yeabsira Asefa Ashengo, Rosa Tsegaye Aga, and Surafel Lemma Abebe. “Context based machine translation with recurrent neural network for English–Amharic translation”. In: *Machine Translation* 35.1 (2021), pp. 19–36.

[76] B Arfaso. “Bi-Directional English-Afan Oromo Machine Translation Using Convolutional Neural Network”. In: *Addis Ababa University* (2019).

[77] Saifu Dereje. “Hybrid Artificial Neural Machine Translation using Deep Learning Techniques English-to-Afaan Oromoo”. PhD thesis. ASTU, 2019.

[78] Gashaw Meron. “Attention-based Amharic-to-Afaan Oromo Neural Machine Translation”. PhD thesis. JU, 2021.

[79] Atnafu Lambebo Tonja, Michael Melese Woldeyohannis, and Mesay Gemeda Yigezu. “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)*. IEEE. 2021, pp. 71–76.

[80] Yitayew Solomon, Million Meshesha, and Wendewesen Endale. “Optimal Alignment for Bi-directional Afaan Oromo-English Statistical Machine Translation”. In: *vol* 3 (2017), pp. 73–77.

[81] Alexis Conneau and Guillaume Lample. “Cross-lingual language model pretraining”. In: *Advances in neural information processing systems* 32 (2019).

[82] Cheonbok Park et al. “Unsupervised neural machine translation for low-resource domains via meta-learning”. In: *arXiv preprint arXiv:2010.09046* (2020).

[83] Guillaume Lample and Alexis Conneau. “Cross-lingual language model pretraining”. In: *arXiv preprint arXiv:1901.07291* (2019).

[84] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. “Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages”. In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 116–126. URL: https://aclanthology.org/2021.mrl-1.11.

[85] Junyoung Chung et al. “Empirical evaluation of gated recurrent neural networks on sequence modeling”. In: *arXiv preprint arXiv:1412.3555* (2014).

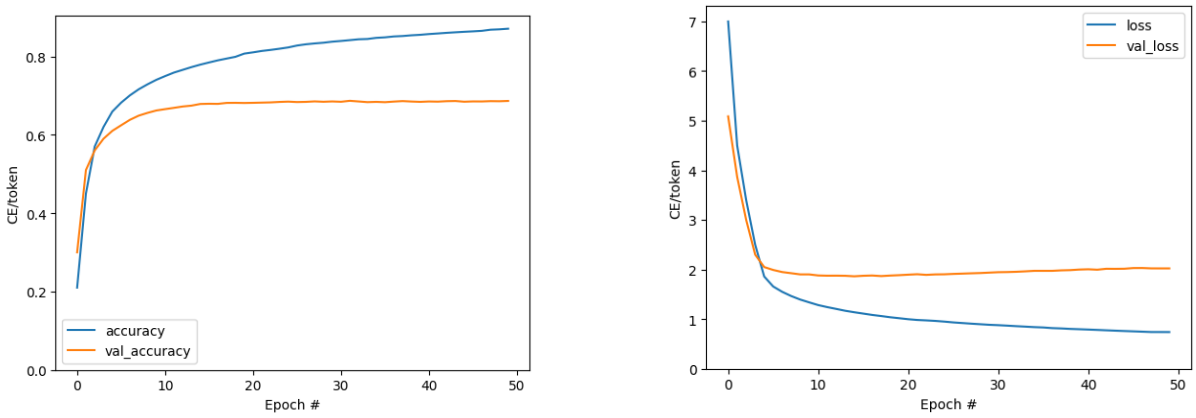
[86] Marta R Costa-Jussa and José AR Fonollosa. “Character-based neural machine translation”. In: *arXiv preprint arXiv:1603.00810* (2016).

[87] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[88] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[89] Robbel Habtamu and Beakal Gizachew. “State-of-the-Art Approaches to Word Sense Disambiguation: A Multilingual Investigation”. In: *Pan African Conference on Artificial Intelligence*. Springer. 2023, pp. 176–202.

Appendix



(a) (b)

Figure 15. *Learning curves of proposed model 15a Learning Accuracy and 15b Loss.*

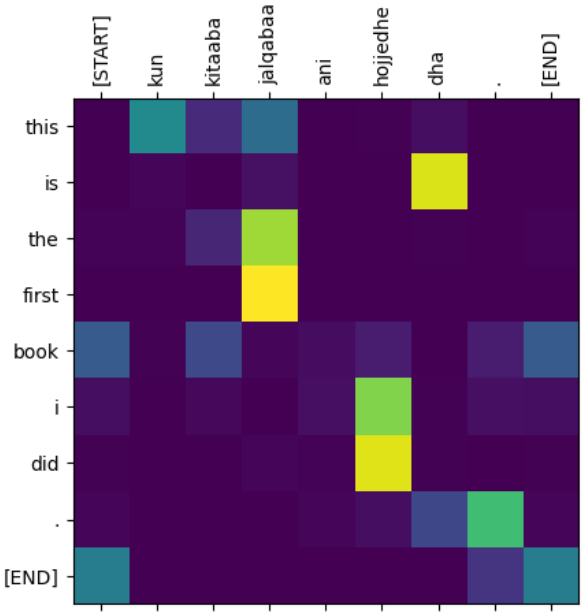


Figure 16. *Sample Attention plot of predicted sentence.*