



JIMMA UNIVERSITY
JIMMA INSTITUTE OF TECHNOLOGY
FACULTY OF COMPUTING

**MORPHOLOGICAL SEGMENTATION USING NEURAL
NETWORKS FOR AFAAN OROMO**

By: Rebuma Regasa

A Thesis Submitted to the Department of Information Technology in Partial
Fulfillment for the Degree of Master of Science in Information Technology

Jimma, Ethiopia
January 2020


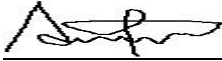
Jimma University
Jimma Institute of Technology
Faculty of Computing

**MORPHOLOGICAL SEGMENTATION USING NEURAL
NETWORKS FOR AFAAN OROMO**

By: Rebuma Regasa

This is to certify that the thesis prepared by **Rebuma Regasa**, entitled **Morphological Segmentation Using Neural Networks For Afaan Oromo**, Submitted in partial fulfillment of the requirements for the Degree of Master of Science in *Information Technology* compiles with the regulations of the University and meets the accepted standards with respect to originality and quality.

Approved by board of Examining Committee:

	Name	Signature
Dean, Faculty of computing:	Getachew Mamo(PhD)	_____
Advisor:	Teklu Urgessa (PhD)	_____ 
External Examiner:	Kula Kekeba(PhD)	_____ 
Internal Examiner:	Getachew Mamo(PhD)	_____

Jimma, Ethiopia

January 2020

Dedicated to

- 1. Regasa Sehata (MY Father)*
- 2. Almaz Nagesso (MY Mother)*

STATEMENT OF THE AUTHOR

As author of this research study, I declare that the thesis is a result of my genuine work, support of my supervisors and help hands of other individuals. Thus, all those had who participated in the study and sources of materials used for writing this thesis have been duly acknowledged. I have submitted this thesis to Jimma University as a partial fulfillment for the requirements of Degree of Master of Science in Information Technology. The library directorate of Jimma University can deposit the copy of the thesis in the university library so that students and researchers can refer it. Moreover, I declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate and/or to get prove of society's problems. Any brief quotations from this thesis are allowed without requiring special permission if an accurate acknowledgement and citation (after publication) of the source is made. In all other instances, however, permission must be obtained from the author.

Rebuma Regasa

Name

Signature

Date

ACKNOWLEDGMENT

First, and foremost glory to Almighty God for giving me full of health, peace, the opportunity to pursue my graduate study at Jimma University and for being with me in all aspects elsewhere. I would like gratefully and sincerely to thank my thesis advisor, **Teklu Urgessa (PhD)**, for his invaluable comments, suggestions, and patience during the entire time of the study. I would like to express my heartfelt thanks to my co-advisor, **Mr Desalew Yohannes(MSc)** for his encouragement, supervision and constructive comments that develop my research abilities and in improving the quality of the Thesis. My special appreciation and acknowledgment go to Mettu University for offering me then opportunity and financial support during my study at Jimma University. Finally, I take this opportunity to sincerely express my gratitude to my Beloved family who are the source of pride and encouragement throughout my life. I am thankful to my father, Regasa Sehata, my mother, Almaz Nageso and to all my brothers and sisters for their encouragement and pray to the success of this work.

Table of Contents

Contents	page
STATEMENT OF THE AUTHOR	iv
ACKNOWLEDGMENT	v
LIST OF ABBREVIATIONS AND ACRONYMS	xi
ABSTRACT	xii
CHAPTER ONE	1
1. INTRODUCTION	1
1.1 Background	1
1.2 Statement of the problem	4
1.3 Research Questions.....	6
1.4 Objective of the Research Work	6
1.4.1 General Objective.....	6
1.4.2 Specific Objectives.....	6
1.3 Methodology of the study	7
1.3.1 Literature review	7
1.3.2 Corpus preparation and development	7
1.3.3 Development Environment and Tools.....	7
1.4 Scope of the Study	8
1.5 Limitation of the study	8
1.6 Application of the results.....	8
1.7 Thesis Organization	8
CHAPTER TWO	9
2. LITERATURE REVIEW	9
2.1. Introduction	9
2.2. Basic Concepts of Morphology	9
2.2.1. Morphemes.....	10
2.2.2 Morphotactics	10
2.2.3 Morphological Processes	11
2.2.4 Computational Morphology.....	13
2.2.5 Morphological Segmentation.....	14

2.3.	Artificial Neural Networks	16
2.4.	Distributed Representation of words.....	17
2.4.1.	Skip-gram model	18
2.4.2.	Continuous Bag of Words (CBOW).....	19
2.5.	Deep learning.....	20
2.5.1.	Recurrent neural networks	20
2.5.2.	Long short-term memory neural networks (LSTM).....	21
2.5.3.	Bidirectional Long short-term memory neural networks (BLSTM)	24
2.5.4	Encoder-Decoder Sequence-to-Sequence Architectures	24
2.6	Summary.....	24
CHAPTER THREE		25
3.	RELATED WORKS	25
3.1.	System for Morphological Processing of Amharic, Afaan Oromo and Tigrinya (HornMorph) .	25
3.2.	Suffix Sequences Based Morphological Segmentation for Afaan Oromo	26
3.3.	Morphological Segmentation with LSTM neural networks for Tigrigna language	26
3.4.	Morphological Analysis of Ge'ez Verbs Using Memory Based Learning	27
3.5.	Morphological Segmentation with Seq2Seq neural network for Russian language.....	28
3.6	Neural Morphological Analysis: Encoding-Decoding Canonical Segments	29
3.7	Summary.....	30
CHAPTER FOUR		31
4.	OVERVIEW OF AFAAN OROMO	31
4.1	Introduction.....	31
4.2	Afaan Oromo Writing System	31
4.3	Afaan Oromo Morphology.....	32
4.3.1	Types of Morphemes in Afaan Oromo	32
4.4	Morphophonemic Processes	54
4.4.1	Reduplication	55
4.4.2	Deletion	55
4.4.3	Assimilation.....	55
4.4.4	Epenthesis.....	55
5	Summary.....	55
CHAPTER FIVE.....		56

5. METHODOLOGY	56
5.1 Architecture of Afaan Oromo Morphological Segmentation	56
5.2 Algorithm of AOMS	60
5.3 Prediction Architecture	61
5.4 Algorithm to Predict Morphemes of word	61
CHAPTER SIX	62
6. EXPERIMENTS	62
6.1 Data Collection.....	62
6.2 Development tools.....	62
6.2.1 Tensorflow	62
6.2.2 Keras	62
6.2.3 Scikit learn.....	63
6.3 Evaluation metrics.....	63
6.4 Experimental Setup.....	64
6.5 Experimental Scenarios	64
6.6 Training.....	65
6.7 Results	68
6.8 Comparison with Baseline Experiments.....	69
6.9 Discussion	71
CHAPTER SEVEN	72
7. CONCLUSION AND FUTURE WORKS	72
7.1 Conclusion	72
7.2 Contribution of the work.....	73
7.3 Future works	73
APPENDICES	74
References	77

List of Tables

Table 1. Summary of related works.....	30
Table 2 Word and its geminated form in Afaan Oromo	32
Table 3. First category of pluralization suffixes.....	35
Table 4. Afaan Oromo pluralization suffixes summarization	38
Table 5. Afaan Oromo suffixes for singulative indicator	39
Table 6. Afaan Oromo suffixes for gender indicator	39
Table 7. Afaan Oromo suffixes for nominative indicator	40
Table 8. Afaan Oromo suffixes for instrumental indicator	42
Table 9. Afaan Oromo suffixes for locative indicator	43
Table 10. Afaan Oromo suffixes for vocative indicator	44
Table 11. Personal pronoun	47
Table 12. Reflective pronoun.....	48
Table 13. Demonstrative pronoun	48
Table 14. Interrogative pronoun.....	48
Table 15. Plural Forms of Adjectives	49
Table 16. Case Inflection Realized on Adjectives.....	50
Table 17. Aspectual Distinction on Verbs	52
Table 18. Affirmative and Negative Verbs in the Imperative mood	53
Table 19. Summary of Marking Jussive Mood [81].....	54
Table 20. Active and passive voices.....	54
Table 21. Hardware/ Software Specification	64
Table 22. Loss and accuracy of model	67
Table 23. Performance results for deep neural networks	68
Table 24. Correctly Segmented, incorrectly segmented and un-segmented morphemes	68
Table 25. Comparison with baseline experiments.....	70
Table 26. Sample of Dataset	74
Table 27. Afaan Oromo vowels and their sounds (adopted from [67])	75
Table 28. Afaan Oromo consonants and their sounds (adopted from [8]).....	75

List of Figures

Figure 1 Overview of how a morphological segmentation works.....	14
Figure 2. Skip gram model	19
Figure 3. CBOW model.....	19
Figure 4. A recurrent neural network and the unfolding in time of the computation involved in its forward computation.	21
Figure 5. LSTM decision process on what information to throw.....	22
Figure 6. LSTM decision process on what information to store	23
Figure 7. LSTM cell state output	23
Figure 8. Architecture of AOMS.....	57
Figure 9. Prediction architecture	61
Figure 10. Model Accuracy	66
Figure 11 Model loss	67
Figure 12. Screen shot of the morphologically segmented Afaan Oromo words.....	69

LIST OF ABBREVIATIONS AND ACRONYMS

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
ANNs	Artificial Neural Networks
AO	Afaan Oromo
API	Application Programming Interface
AOMS	Afaan Oromo Morphological Segmentation
BLSTM/BiLSTM	Bidirectional Long short-term memory neural networks
CBOW	Continuous Bag of Words
CPU	Central Processing Unit
CV	Consonant Vowel
CVC	Consonant Vowel Consonant
DNNs	Deep Neural Networks
FST	Finite State Transducer
LSTM	Long short-term memory neural networks
MIT	Massachusetts Institute of Technology
NLP	Natural Language Processing
NLU	Natural Language Understanding
OOV	Out-Of-vocabulary
POS	Part of Speech Tagging
RMSprop	Root Mean Square Propagation
RNNs	Recurrent neural networks
Seq2Seq	Sequence to Sequence
SGD	stochastic Gradient Descent

ABSTRACT

Natural Language Processing (NLP) concerns with computational processing of natural languages in order to provide a products as computers interact linguistically with people in ways that suit people rather than computers. Morphological segmentation is one of the applications of natural language processing that studies the use of computer programs and software to segment words to their morphemes. Morphological segmentation is used as components in many applications, specially machine translation, spell-checker, Part of Speech Tagging (POS) tagging. Several researchers have applied machine learning approaches for Afaan Oromo morphological segmentation while no research have used artificial neural networks for morphological segmentation task.

Artificial neural network is subset of machine learning which inspired by the structure, processing method and learning ability of a biological brain. The processing of multiple data inputs is done by different machine learning algorithms. Hence, Neural Networks have the ability to learn by themselves and produce the output that is not limited to the input provided to them.

Morphological segmentation using neural networks have been developed for languages such as English. Thus, the main aim of study is to development of a morphological segmentation using neural networks for Afaan Oromo. In order to achieve the objective of this research work, a corpus is collected from different sources such as Books, Newspapers of Afaan Oromo and prepared in a format suitable for use in the development process. We have used corpus of size 50,200, which we have been developed. From this corpus we have used corpus of size 40,160 for training and 10,040 for testing of our work. From the experiments F-score achieved was 97.48%, 98.33%, 98% using Bidirectional Long Short Term Memory, Long Short Term Memory, and Recurrent Neural Networks respectively.

In conclusion, the accuracy of the Afaan Oromo morphological segmentation using neural networks were promising than baseline experiments. To improve the performance of the model increase number of training data were recommended for future works

Keywords: *Morphology, Morphological Segmentation, Afaan Oromo, Neural Networks*

CHAPTER ONE

1. INTRODUCTION

This chapter discusses about basic concept of morphology, the objective, problem statement, methodology, and application of the study.

1.1 Background

Language is one of the fundamental aspects of human behavior and it constitutes a crucial component of our lives [1, 27]. In its written form it serves as a means of recording information and knowledge on a long term-basis and transmitting what it records from one generation to the next. In its spoken form, it serves as a means of coordinating our day-to-day life with others [1].

Linguistics are the study of natural language; it involves analyzing language form, language meaning, and language in context. Natural language is a language that involved as a means of communication among people. An approach to linguistics that employs methods and techniques` of computer science is called computational linguistics or Natural language processing. Natural language processing has been developed in the 1960s, as a subfield of Artificial intelligence and linguistics [2]. Natural language processing is a field of computer science that investigates interactions between computers and human languages, which is used for both generating human readable information from computer systems and converting human language into more formal structures that a computer can understand [3].The ultimate goal of natural language processing is for computers to achieve human-like comprehension of languages. To achieve this, the machine should understand first the natural language before processing it. Thus, Natural language processing demands deep natural language understanding and modeling the natural language so that computer programs that act appropriately on the information contained in the text can be developed. When this is achieved, computer systems will be able to understand, draw inferences from, summarize, translate, and generate accurate and natural human text and language.

While Natural language processing is a relatively recent area of research and application, as compared to other information technology approaches, there have been sufficient successes to date that suggest that Natural language processing-based information access technologies will continue to be a major area of research and development in information systems now and far into the future

[4]. Well known research areas of NLP are morphological segmentation, part of speech tagging, word sense disambiguation, and machine translation [3].

There are different levels of natural language processing, for instance, systems developed for processing NL at phoneme, word, sentence, and pragmatic levels [27]. These systems are developed in such a way that the output of a lower system can serve as an input to the next higher level. Therefore morphological segmentation can be used in many Natural Language Processing applications such as machine translation, speech recognition, and part-of-speech tagging [12]. This study is based on the morphological level of natural language processing

Morphology is the branch of linguistics and one of the major components of grammar that studies word structures, especially in terms of morphemes, which are the smallest units of language [5]. It deals with the inner structure of individual words and the laws concerning the formation of new words from pieces, morphs. There are two types of morphemes. Free morphemes and bound morphemes. Free morphemes can stand alone with a specific meaning, for example, eat, date, weak. Bound morphemes cannot stand alone with meaning. E.g. for Afaan Oromo word ‘Sangoota’, Sang- is free morpheme, and the plural suffix -oota is a bound morpheme. Morphemes are comprised of two separate classes called (a) bases and (b) affixes. A base or root is a morpheme in a word that gives the word its basic meaning. An affixes are a bound morpheme that occurs before or after a base.

Bound morphemes are divided into two types, inflectional and derivational morphemes. Derivational morphemes are used to make new words in the language and are often used to make words of a different grammatical category from the stem [6]. *Ex. good => adjective good + ness => noun Care => noun care + less => adjective and from Afaan Oromo bar- ‘to know’ (v) barumsa ‘education’(bar-+-umsa)(noun). Inflectional morphemes are not used to produce new words in the English language, but rather to indicate aspects of the grammatical function of a word. Plural markers, possessive markers, tense markers, comparative and superlative markers are inflectional morphemes. Ex. for Afaan Oromo saree- ‘dog’ and saree + -oota (bound morpheme which indicates pluralization) = saroota ‘dogs’.*

Morphological segmentation is the task of detecting morphological boundaries. This task involves the breaking down of words into their component morphemes. For example, the English word “reads” can be segmented into “read” and “s”, where “read” is the stem and “-s” is an inflectional morpheme, marking third person singular verb. Morphological segmentation is useful for several downstream NLP tasks, such as morphological analysis, POS tagging, and Machine Translation [12]. Word segmentation is always applied as preprocessing phase in many NLP applications, such as machine translation, information retrieval, and speech recognition.

Morphological segmentation can be implemented with many machine learning algorithms. Neural network or in modern sense an artificial neural network is type of the machine learning, which is inspired by the structure of the brain and it contains highly interconnected entities called units or nodes. A neural network is an interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the inter unit connection strengths, or weights, obtained by a process of adaptation to, or learning from, a set of training patterns [7]. A neural network is group of algorithms; these algorithms model the data using neurons for machine learning.

Afaan Oromo is one of the major African languages that are widely spoken in Ethiopia. Oromo people are the native speaker of this language; they are largest ethnic group in Ethiopia as well as in east Africa. According to the Ethiopia Population Census Commission 2007 the total population of Oromo People is 34.5% of the total population of the country [9]. Afaan Oromo is also spoken in the neighboring countries like Kenya, Uganda, Tanzania, Djibouti and Somalia. Currently, it is an official language of Oromia Regional State. With regard to the writing system, Qubee(a Latin-based alphabet) has been adopted and become the official script of Afaan Oromo since 1991.

As noted in [13], although resource-rich languages such as English have well-developed language tools, low-resource languages suffer from either low grade or the absence of electronic data support altogether to pursue NLP research. Afaan Oromo language is one such low-resource language and has been largely ignored in NLP-related research, mainly due to the absence of a text corpus. Therefore, we based our studies on a new morphologically segmented corpus developed in-house. In this paper, we focus on the task of designing and implementing morphological segmentation using neural networks for Afaan Oromo.

1.2 Statement of the problem

There are more than 80 languages in Ethiopia. Afaan Oromo (when translated it means *Oromo Language*) is one of the major Languages that is widely spoken and used in Ethiopia [8]. As the Ethiopia's statistical report of 2007 [9] shows there are more than 25 million speakers of Afaan Oromo in Ethiopia and this fact shows that, the language has the largest speaker in the Country. Afaan Oromo has 26 basic symbols. It is academic language of primary schools of Oromia region in Ethiopia and official language of the region. Oromo language, literature and folklore delivered as a field of study in many universities located in Ethiopia and other countries [10]. Nowadays journal, magazines, newspapers, news, online education, books, entertainment Medias, videos ,pictures, are available in electronic format both on the Internet and on offline sources [11].

Developing complex natural language processing applications such as machine translation, speech recognition, POS tagging, word sense disambiguation with high performance requires preprocessing phase such as morphological segmentation. As mentioned above, morphological segmentation is task of breaking down of words into their component morphemes. Many researches have been studied morphological segmentation for different languages. These researches used different approaches, such as Sirts et al [16] worked on minimally supervised morphological segmentation using adaptor grammars. Adaptor Grammars (AGs) are a nonparametric Bayesian modelling framework that can learn latent tree structures over an input corpus of strings. Adaptor grammars define morphological grammar where each word consists of zero or more prefix, stem and zero or more suffixes. Morphological segmentation using sequence to sequence (seq2seq) using neural network model was developed by [17]. Seq2seq model is consisting of an encoder, decoder. An encoder reads source data and produces a vector containing information about this data relevant to the task. Decoder learns to generate a translation of encoded data. Yemane [18] tried to develop morphological segmentation with LSTM(Long-Short Term Memory) neural networks for Tigrigna. Yemane employed Conditional Random fields and window based long-short term memory neural networks separately to develop boundary detection models. Wang [19] Proposed novel neural network architecture which relay on long-short term memory neural networks. It learns the structure of input sequences from raw input words and then it can predict morphological boundaries.

As far as researcher's seen very limited works have been done for Afaan Oromo morphological segmentation such as, HornMorph [15] attempted to develop morphological analysis for three languages; Afaan Oromo, Amharic and Tigrigna using finite state transducers and Getachew [12] demonstrated morphological segmentation for Afaan Oromo using machine learning approach. [15] Focuses on morphological features of nouns and verbs of the languages. Hornmorph can identify root part of word. But, limitation of hornmorph were; it mainly explains words based on their morphological features without segmenting into their all morpheme components. As an example: '**dubbadhu**'/speak/. It takes this words and display: 'dubbadh-'. The removed part from this words was the first morph only. But, in reality of Afaan Oromo grammar this word and like this words have more than one morphemes (**dubbadhu: dubb-adh-u**). In this case **dubb-** can be root of the many words such as dubb-at-e, dubb-ii, dubb-ach-uu...etc, **-adh-** indicates second person. Again **dubbadh-** can be root of other words such as dubbadh-e, dubbadh-u, dubbadh-a.

The second researched on this area [12] was demonstrated morphological segmentation for Afaan Oromo using machine learning approach which learns suffix sequences of word. Suffix sequences are set of suffixes that attached to the root of word which can allow us to identify boundaries between morphemes of many words in Afaan Oromo. Under this research almost words of Afaan Oromo languages are segmented into roots and each morphemes. In [12] work, some of words are not segmented due to missing suffix sequences of words. As an example: **Kenne**/He gave. Their model was segmented into: **ken-n-e** instead of **kenn-e**. In this case **ken-n-e** is not correct segmentation for word **kenne** as indicated by rule of Afaan Oromo morphological segmentation. On the other hand **kenn-e** is correct and it can be root of other words such as kenn-e, kenn-u, kenn-at-a and other words.

Morphological segmentation can be performed using rule-based, machine learning approaches or deep learning. Rule-based approaches can be quite expensive and language dependent because the morphemes and all the affixation rules need to be identified to disambiguate segmentation boundaries. Machine learning approach, on the other hand, is data-driven wherein the underlying structure is automatically extracted from the data. Deep learning is subset of machine learning where algorithms are created and function similar to those in machine learning. The primary task in developing a machine learning based morphological segmentation is to learn the rules automatically which is the toughest task since the complexity of feature representation is high.

Therefore, in this research, we proposed a neural network approach for learning the rules of morpheme separation automatically.

1.3 Research Questions

The following research questions will be answered in the research:

1. How to develop morphological segmentation model using neural network for Afaan Oromo?
2. How to improve performance of morphological segmentation for Afaan Oromo using neural networks?
3. How to determine and customize deep learning parameters to improvement the performance of the segmentation?

1.4 Objective of the Research Work

1.4.1 General Objective

The general objective of the research was to investigate the use of neural network approaches to the development morphological segmentation for Afaan Oromo.

1.4.2 Specific Objectives

In order to achieve the general objective stated above, the study attempts the following specific objectives:

- ✓ To review, analyze and understand the theoretical basis of morphological segmentation and identify the relations that exist between the previous research attempts in the area.
- ✓ Studying morphological properties of Afaan Oromo words
- ✓ To collect and design corpus for training and testing of the model
- ✓ Building deep neural network models that use word vector features as input
- ✓ Evaluating performance of new developed models
- ✓ To forward conclusion and recommendation based on the findings

1.3 Methodology of the study

The methodology part focuses on the approaches, techniques, tools, and data source the researcher plans to follow during the course of the study. Methodology provides an understanding of how a proposed research was conducted in order to obtain information for developing the proposed systems [14]. This section is very necessary as the performance of the final system depends on reasonable approaches and techniques followed by the researcher. Accordingly, the tools, study design and others that the researcher plans to follow are detailed below.

1.3.1 Literature review

For better understanding of the area, a number of related works and resources which consists of the journal articles; books and internet have been reviewed. Detailed investigation of different literatures in different languages helped the researchers to get an insight of the existing problems. Furthermore it would help the researcher to contextualize the techniques and methodologies used by different scholars and help to come up with the suitable approach for the problem under study. In this study review is mainly concerned works that have direct relation with the topic and the objective of the study.

1.3.2 Corpus preparation and development

There is no publicly available morphologically segmented resource for Afaan Oromo. As the result, we collected electronic data from Afaan Oromo books, newspapers and dictionaries with the help of instructors of Afaan Oromo department in Jimma University. The current version of this corpus comprises over 50,200 tokens with it's with target training corpus (morphemes). Every word in training corpus is represented as the sequence of its letters, e.g **deemeera** (he is gone). The special symbol is “-” was added in target training corpus (morphemes). This symbol indicated the boundaries between word's segments, e.g deem-e-era.

1.3.3 Development Environment and Tools

To develop morphological segmentation with neural networks for Afaan Oromo we used windows operating system and python 3.7 programming language.

1.4 Scope of the Study

The scope of the research is limited to exploring neural network approaches to design morphological segmentation for Afaan Oromo. The study was also focused on collecting of words with their morphemes. Compound word segmentation is out of the scope of this research due to the absence of clearly stated rules in literatures.

1.5 Limitation of the study

Main limitation of the research is lack of readily available morphologically segmented corpus of the language. Therefore, we based our studies on a new morphologically segmented corpus developed in-house. As a result, limited corpus was used for evaluating the performance of the system developed in the study. This is because of it takes more time to prepare morphologically segmented words. The absence of well-studied linguistic materials presented various challenges to segment several words into their morphemes.

1.6 Application of the results

Morphological segmentation is considered as important component in many NLP applications. For most natural language processing applications it act as preprocessing phase. Morphological segmentation is serves as an important preprocessing tool and improves the performance for tasks such as machine translation [21] application, in which morphological information is needed to analyze and generate words. Also, [12] identifies morphological segmentation was important component used to improve performance of text retrieval. Based on these facts, Morphological segmentation for Afaan Oromo plays a crucial role in information extraction based researches and applications associated to the language.

1.7 Thesis Organization

This thesis paper is organized into Seven Chapters including the current one. Chapter Two presents literature review on morphological segmentation. Chapter Three presents different related works on morphological segmentation. Chapter Four presents overview Afaan Oromo. The proposed work and experiment results are discussed in Chapter Five and chapter six respectively. The conclusion and future works are presented in Chapter Seven.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Introduction

In this chapter, we briefly discuss overview of morphological segmentation development strategies and approaches. As discussed in chapter 1, the main objective of the study is to develop morphological segmentation for Afaan Oromo language. Morphology is the study of internal structure of the word. Different components of morphological segmentation are reviewed and presented. Morphological segmentation is basic process in any natural language processing task. It is the process of segmenting a given word into its sequences morphemes. Morphological segmentation is the task detecting morpheme boundaries [22]. Morphological segmenter is computer program which receives a word as input and gives morphemes of a word as output.

2.2. Basic Concepts of Morphology

The term morphology is generally attributed to the German poet, novelist, playwright, and philosopher Johann Wolfgang von Goethe (1749–1832), who coined it early in the nineteenth century in a biological context. The term morphology is Greek and is makeup of word ‘morph-’ which means ‘shape, form’ and ‘-ology’ which means ‘the study of something’. The meaning of morphology is various in different disciplines. In biology morphology refers to the study of size, shape and structure of plants, animals and microorganisms and of their relationships of constituent parts. In geology it refers to study of structure of rocks and landforms. In linguistics morphology is the study of words, how they are formed and their relationship to the other words in the same language. It analyzes the structure of words and part of words such as stems, root words, prefixes and suffixes. The term morphology was first used in linguistics by August Schleicher in 1859. Linguistics distinguish simple words, such as ‘soon’ which has no internal structure apart from sound, and complex words such as, ‘sooner’ which can be analyzed into meaningful parts(in this case ‘soon’ and the English comparative suffix ‘-er’). In modern-day linguistics, which began in the 19th century, morphology is one of the core areas of grammar, along with phonetics, phonology, syntax, and semantics/pragmatics.

2.2.1. Morphemes

A major way in which morphologists investigate words, their internal structure, and how they are formed is through identification and study of morphemes. A morpheme is the smallest part of a word that has grammatical function or meaning but not the smallest unit of meaning. For example, *sawed*, *sawn*, *sawing*, and *saws* can all be analyzed into the morphemes {saw} + {-ed}, {-n}, {-ing}, and {-s}, respectively. None of these last four can be further divided into meaningful units and each occurs in many other words, such as *looked*, *mown*, *coughing*, *bakes*. {Saw} can occur on its own as a word; it does not have to be attached to another morpheme. It is called free morpheme. However, none of the other morphemes listed just above is free. Each must be affixed (attached) to some other unit; each can only occur as a part of a word. Morphemes that must be attached as word parts are said to be bound morpheme. Affixes can be prefixes, suffixes, circumfixes and reduplication [27]. Prefixes are attached before base. In English /re-/, /un-/, /im-/ in *rewrite*, *unable*, *immoral* are prefixes. In Afaan Oromo /wal-/, /al-/ in *wal-faana* (together) and *al-tokko* (once) are prefixes. On the other hand suffixes are attached after base. In English /-s/, /-ed/, /-ing/ in *cats*, *walked*, *talking* are suffixes. In Afaan Oromo /-oota/, /-een/ in *sangoota* (Oxen), *manneen* (houses) are suffixes. *Circumfixes* are affixes that surround the word, attaching to the beginning and end of the word. In English the combination of the prefix ‘un-’ and suffix ‘-ed’ as in the case of *unnoticed* is an example of circumfix. There is no circumfix in Afaan Oromo. Reduplication refers to words formed through repetition of sounds. Reduplication is used in Afaan Oromo to show an action done repeatedly [23]. For example, in Afaan Oromo *kute* (he cut) and *kukkute* (he cut into pieces).

2.2.2 Morphotactics

Every known language has a kind of rules that govern the arrangement of morphemes to create a word. Usually, it is word grammar that determines the way morphemes put together to form words. A subfield of morphology that deals with such rules is called morphotactics. Usually there are language specific words grammars that help determine how morphemes are put together. These word grammars put constraints on morph patterning. For example, the English word *pseudohospitalization* is formed from /pseudo-/, /hospital/, /-ize/ and /-ation/. But these morphemes can be concatenated randomly as follows if such word grammars don’t restrict their formation: *hospitalationizepseudo, pseudoizehospitalation, pseudohospitalationize*[24, 25].

For example, in Afaan Oromo from morphemes ‘nyaat-’ (to eat), /-anii/ and /-ru/ different word forms can be generated, as in /runyaatani/, /nyaatruanii/ and /nyaataniiru/, but the grammatically correct one is ‘nyaataniiru’ (they have eaten) because the suffixes follow the stem in Afaan Oromo. In the order for morphological analyzer to parse surface complex words robustly back to atomic constituents, it should take the morphotactic rules into account.

Besides to the morphotactic rules, phonological rules are also applied during the formation of a word from morphemes which, then, results in phonological and assimilation effects. For example: when the English prefix /in-/ concatenates with a free morpheme /possible/, then it becomes influenced (changed into /im-/) and produce the complex word /impossible/ as the result of the concatenation process. A discipline called Morpho-phonology, which merges both morphology and phonology, deals with these changes and their fundamental reasons. When morphemes concatenate to form a larger unit, the sound or shape of morphemes may be influenced and results in orthographical or phonological changes. This kind of influence is called phonological influence. Phonology is another subfield of linguistics that study the structure and systematic patterning of sounds in human language. Indeed, the morphological analyzer should also have a component which takes care of these phonological changes encountered while parsing the complex words [26].

2.2.3 Morphological Processes

There are two major ways to form words from morphemes: inflection and derivation [26].

2.2.3.1 Derivational Morphology

Derivational morphology is concerned with forming new lexemes, that is, words that differ either in syntactic category (part of speech) or in meaning from their bases [26, 27]. Derivation is typically contrasted with inflection, which is the modification of words to fit into different grammatical contexts. Derivational morphemes are added to forms to create separate words: {-er} is a derivational suffix whose addition turns a verb into a noun, usually meaning the person or thing that performs the action denoted by the verb. For example, {paint} + {-er} creates painter, one of whose meanings is “someone who paints” and in Afaan Oromo ‘**baruu(V)**’ and ‘**barnoota(N)**’.

2.2.3.2 Inflectional Morphology

Inflectional morphology is the study of the processes (such as affixation and vowel change) that distinguish the forms of words in certain grammatical categories [26]. Inflectional morphemes do not create separate words. They merely modify the word in which they occur in order to indicate grammatical properties such as plurality, as the {-s} of magazines does, or past tense, as the {ed} of painted does.

2.2.3.3 Compounding

In linguistics, a compound is a lexeme consists of more than one stem. Compounding or composition is the process of word formation that creates compound lexemes. Every language follows certain rules by which it forms its compound [27]. So, it is not the case that all words can combine to form compound words. For example, in Afaan Oromo ‘miila’ means foot and ‘jala’ means ‘under’ and when it combine together ‘miiljala’ means underfoot. Another way that words derived by compounding differ from words derived by affixation is that a compound word doesn’t really have a base or root that determines the meaning of the word. Instead, both pieces of a compound make a sizeable contribution to the meaning. On the basis of the structural changes of the root and other morphemes during affixation, morphology can be classified as linear or nonlinear [28]. Prefixes and suffixes are linear morphology. In linear morphology affixes are added to the root without changing the internal structure of the root, though some euphonic changes might take place at the boundary (of the affixes and the root). On the other hand, morphological systems where the internal structure of the morphemes changes during the addition of affixes are classified as nonlinear morphology [29]. Pluralization and adjectivization in English [30] normally pertains to linear morphology whereas Semitic languages [29] features nonlinearity. Morphological processes in Afaan Oromo are mostly linear in nature [27].

2.2.4 Computational Morphology

Computational morphology is sub-field of computational linguistic that deals with computational analysis and synthesis of word forms and their developing theories [31]. By computational analysis of morphology, one can extract any information encoded in a word and bring it out so that later layers of processing can make use of it [32]. It performs morphological tasks with the help of computers and computational methods automatically. The purpose of such work is to aid in the effective means of the storage of words in lexicons, and provide time-efficient lookup capabilities. Two kinds of processing are of interest: morphological analysis, by which a surface word form is analyzed into a lexical representation, consisting of the word's component morphemes or grammatical features, and morphological generation, by which a lexical representation is converted to a surface word form [33]. Generally, the tasks involved in computational morphology can be grouped into two parts: *Word-form synthesis and analysis*; and *Parts-of-speech (POS)-or inflectional-category determination*. The following section discusses these tasks in detail.

2.2.4.1 *Word-form synthesis and analysis*

Morphological synthesis or generation is a process of returning one or more surface forms from a sequence of morpheme glosses, whereas analysis or recognition does the reverse process, which is tokenizing word forms into their ingredient morphemes. A word form synthesis/ generator would accept as input a lexical form (such as /read/ + /-ing/) and returns surface form /reading/. On the other hand analyzer accept surface form (such as /reading/) and returns its lexical form namely, /read/ + /-ing/. These processes demand identification of word form components (for example stems and suffixes) and taking account of the regular phonological or orthographical alternations due to morphological, and morpho-phonological processes involved [26].

2.2.4.2 *Parts-of-speech (POS)-or inflectional-category determination*

Part of Speech (POS) Tagging is the essential basis of Natural Language Processing (NLP). It is the process in which each word is assigned to a corresponding POS tag that describes how this word is used in a sentence. Computational morphological systems operate as a morphological front end of syntactic parsers. Syntactic parser determines the POS or inflectional category of the entire

word. The POS tag or inflectional category of word is often taken from a morpheme that serves as a “head of a word [26]. For example, the English word 'goodness' comes from the following morphemes: the adjective stem /good/ and the noun marker suffix /-ness/. Of these morphemes, the noun-marker suffix /-ness/ is a head of the word /goodness/, and thus, the entire word is a noun.

2.2.5 Morphological Segmentation

Morphological segmentation is the process of segmenting words into its morphemes and analyzing word formation [34]. The computer program that used for word segmentation is used in other applications such as Speech Synthesis, Search Engines, Grammar Checker and Machine Translation. Afaan Oromo is morphologically rich language and belongs to Latin language family. The process of segmentation can be defined as: for given word W, morphological analyzer returns a list of morphemes M1, M2,...Mn. For example, consider English word ‘reading’. Here ‘W’ is ‘reading’ and its morphemes are ‘read’ and ‘-ing’.

2.2.5.1 How a Morphological Segmentation works

The morphological analyzer starts its segmentation by taking surface forms of words as an input and returns decomposition of words into parts called morphemes. The following figure shows the input and output of a morphological analyzer.

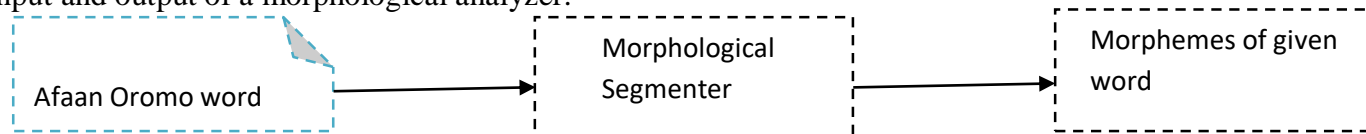


Figure 1 Overview of how a morphological segmentation works

As discussed so far, morphology is all about structure of words. Hence, words of Afaan Oromo are fed into the segmenter. Then, segmenter returns sequences of morphemes of the word. For example, if the input to the segmenter is a word **argameera**/founded/ then the segmenter divide the given word in to following morphemes of the word.

argameera -----arg-am-e-era

2.2.5.2 Approaches to Morphological Segmentation

At this point we turn our attention to what are the approaches actually applied for morphological segmentation. Before discussing in detail, we point out there two basic approaches of morphological segmentation which we call as: rule based approach, corpus based approach. As

[35] demonstrated the different approaches to morphology are categorized as corpus based and rule-based.

- ✓ Rule Based Approaches
- ✓ Corpus Based Approaches

2.2.5.2.1 Rule Based Approaches

Rule-based approaches are based on a theory of morphology laid down by experts. This group of methods enables one to incorporate sophisticated linguistic theory, such as generative phonology, into computational morphology processes. Rule based approaches rely on handcrafted language rules prepared by language experts [36]. But the main disadvantages of this approach are its lack of portability, robustness, and high cost of maintenance in slight change of data [37].

2.2.5.2.2 Corpus based Approaches

Modern natural language processing (NLP) applications, such as speech recognition, information retrieval, and machine translation, perform their tasks using corpus based approaches [38]. It is also called machine learning approaches; do not strictly follow explicit rules of linguistics. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves [39]. Approaches in this category use some algorithms to learn. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Machine learning algorithms are often categorized as *supervised* or *unsupervised* [40]

2.2.5.2.2.1 Unsupervised Machine Learning Algorithms

These algorithms autonomously discover morpheme segmentations in unannotated text corpora [34, 39]. They are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from

unlabeled data. Unsupervised approaches are attractive due to the availability of large quantities of unlabeled text.

2.2.5.2.2.2 Supervised Machine Learning Algorithms

It can apply what has been learned in the past to new data using labeled examples to predict future events [38]. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. Additionally, there is also semi-supervised machine learning approach, in which make use of unlabeled data for training typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning and supervised learning.

The primary task in developing a machine learning based model is to learn the rules automatically which is the toughest task since it requires a huge amount of data require careful feature engineering [41,42]. In this paper, we present a model of morphological segmentation based on a neural network called encoder-decoder model for learning the rules of morpheme separation automatically. It is also called as neural sequence-to-sequence (seq2seq) models which defined the state of the art for morphological segmentation. The following sections discuss overview of artificial neural networks and deep learning.

2.3. Artificial Neural Networks

Artificial Neural Networks (ANNs) are relatively crude electronic models based on the neural structure of the brain. The brain learns from experience. As [43] defined it as an ‘artificial neural network (ANN) is an imitation of the human brain’. A natural brain has the ability to learn new things, adapt to new and changing environment. The functions of the brain are to analyze incomplete and unclear, fuzzy information, and make its own judgment out of it. Artificial neural networks try to mimic the functioning of brain [44]. Even simple animal brains are capable of functions that are currently impossible for computers. Computers do the things well, but they have trouble recognizing even simple patterns.

Brain is a highly complex organ that controls the entire body. The brain of even the most primitive animal has more capability than the most advanced computer. Its function is not just controlling the physical parts of the body, but also of more complex activities like thinking, visualizing, dreaming, imagining, learning etc, activities that cannot be described in physical terms. An artificial thinking machine is still beyond the capacity of the most advanced supercomputers.

The brain stores information as patterns. Some of these patterns are very complicated and allow us the ability to recognize individual faces from many different angles. This process of storing information as patterns, utilizing those patterns, and then solving the problems encompasses a new field in computing, which does not utilize traditional programming but involves the creation of massively parallel networks and the training of those networks to solve specific problems.

Human brain is made of cells which are most basic element called neurons [43]. These neurons provide the abilities to remember, think, and apply previous experiences to our every action. Interconnection of such cells (neurons) makes up the neural network or the brain [45]. The power of the human mind comes from the numbers of these basic components and the multiple connections between them.

But, artificial neural networks do not approach the complexity of the brain. There are, two key similarities between biological and artificial neural networks. First, the building blocks of both networks are simple computational devices called neurons that are highly interconnected. Second, the connections between neurons determine the function of the network.

2.4. Distributed Representation of words

Neural Networks are designed to learn from numerical data. Word vector really all about improving the ability of networks to learn from text data. Word representation is encoding of raw text to vector of numbers that is convenient for use by machine learning algorithms. Word representation , also known as word embedding, plays an increasingly vital role in building continuous word vectors based on their contexts in a large corpus [46]. Word embedding captures both semantic and syntactic information of words, and can be used to measure word similarities, which are widely used in various IR and NLP tasks. Deep learning systems give each word a distributed representation, i.e., a dense low dimensional real-valued vector or an embedding. The

main advantage of having such a distributed representation over word classes is that it can capture various dimensions of both semantic and syntactic information in a vector where each dimension corresponds to a latent feature of the word [47]. Word representation methods may be broadly classified into frequency or count-based methods which employ co-occurrence of words and prediction-based approaches that assign probabilities to measure degree of relatedness. An elaborate comparison of both methods is discussed by [48]. Several architectures have been proposed for building word embedding and using them as features to improve NLP tasks. [49] Proposed a new distributed representation of words that processes very large datasets with significantly lower computational cost. Word2Vec is one of the most popular technique to learn word embeddings using shallow neural network. We utilized word2vec for training on large text data for representing the words and morphemes as relatively low dimensional dense vectors. It can be obtained using two methods; continuous bag of words (CBOW) and the Skip-gram model. Their difference is that CBOW predicts the target word from the contextual information whereas the Skip-gram model predicts the surrounding words, given the target word. These methods are briefly explained in the following subsections.

2.4.1. Skip-gram model

This model accepts a word W_i and predicts the words around a given word (W_i), which are context words (W_{i-2} , W_{i-1} , W_{i+1} , W_{i+2}). A context word does not need to be immediate words. Some words can be skipped within a given window size to look forward and backward from target word. Skip gram model has one hidden layered neural network. The input layer consists of one-hot encoded vector of the vocabulary. Skip gram model neural network is depicted in figure below

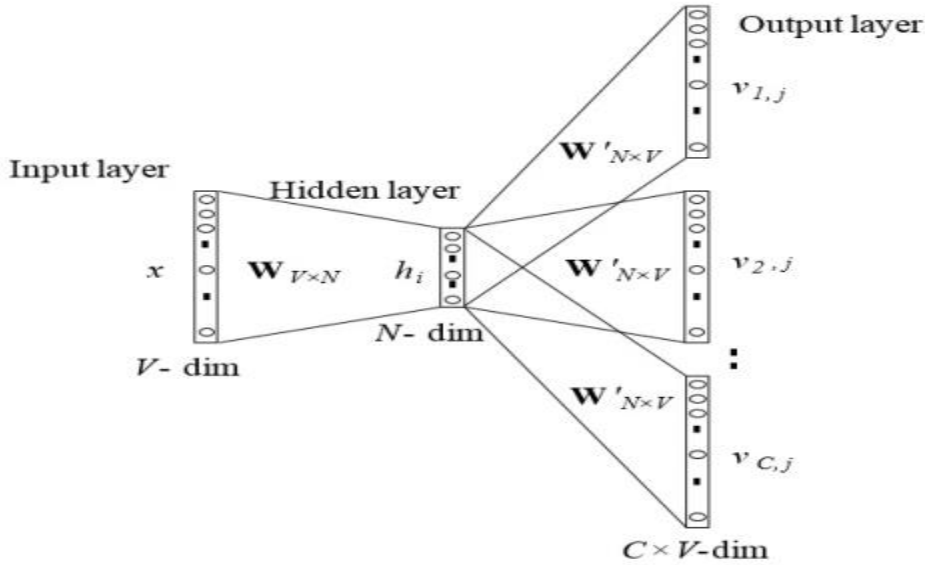


Figure 2. Skip gram model

2.4.2. Continuous Bag of Words (CBOW)

Continuous bag of words is reverse of skip gram model. Given the context (W_{i-2} , W_{i-1} , W_{i+1} , W_{i+2}) the task is to predict the word. Continuous bag of words model (CBOW) takes the average of the vectors of the input context words to compute the output of hidden layer, and use the product of the input layer hidden layer weight matrix and the average vector as the output. CBOW model neural network is depicted on Figure below

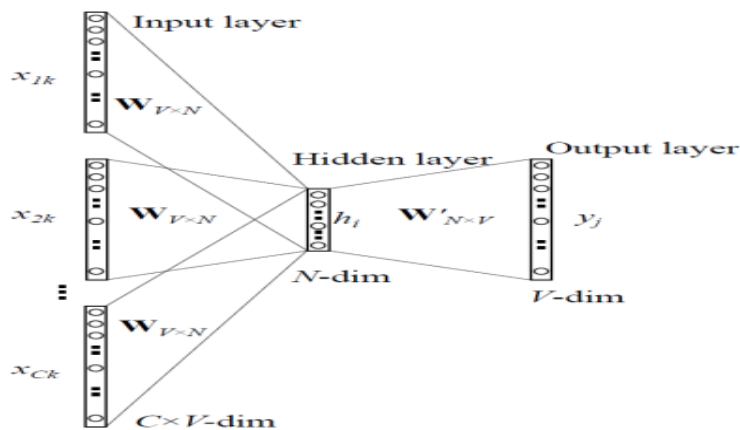


Figure 3. CBOW model

2.5. Deep learning

Since 2006, deep structured learning, or more commonly called deep learning or hierarchical learning, has emerged as a new area of machine learning research. During the past several years, the techniques developed from deep learning research have already been impacting a wide range of signal and information processing work within the traditional and the new, widened scopes including key aspects of machine learning and artificial intelligence [50].

Deep learning is a set of machine learning algorithms that attempt to learn layered model of inputs commonly known as neural nets. It allows computational models composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in many NLP applications [42].

Various deep neural network architectures such as convolutional deep neural networks, deep belief networks and recurrent neural networks have been applied to many areas and shown impressive results. These include natural language processing computer vision speech recognition and bioinformatics [51]. There are many types of deep neural networks and deep neural network architectures. This paper briefly discusses neural networks used in our research.

2.5.1. Recurrent neural networks

Traditional neural networks have a major limitation in considering sequential relation of inputs and outputs. It is assumed that each inputs and outputs are independent of each other. To overcome this limitation recurrent neural networks (RNN) are proposed. Recurrent neural networks (RNNs) have been widely used for processing sequential data [52].

Recurrent neural networks have memory about what has been calculated so far and uses it on current output computation. Theoretically RNN make use of information in arbitrarily long sequence but in practice it is limited to few time steps [53]. Typical recurrent neural network is shown in Figure

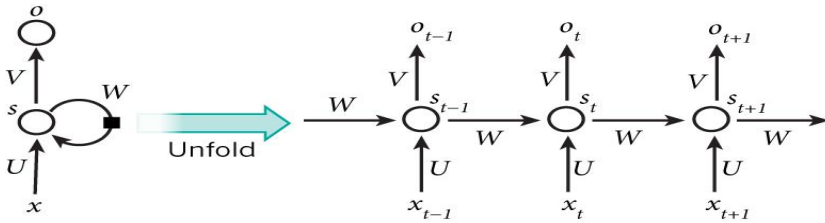


Figure 4. A recurrent neural network and the unfolding in time of the computation involved in its forward computation.

The formulas that govern the computation happening in a RNN are as follows:

- ❖ X_t is the input at time step t . For example, X_1 could be a one-hot vector corresponding to the second word of a sentence.
- ❖ S_t is the hidden state at time step t . It's the memory of the network. S_t is calculated based on the previous hidden state and the input at the current step: $S_t = f(Ux_t + Ws_{t-1})$. The function f usually is a non-linearity such as [tanh](#) . s_{t-1} which is required to calculate the first hidden state, is typically initialized to all zeroes.
- ❖ o_t is the output at step t .

Unlike a traditional deep neural network, which uses different parameters at each layer, a RNN shares the same parameters (U , V , and W above) across all steps. This reflects that it is performing the same task at each step, just with different inputs. This greatly reduces the total number of parameters the network needs to learn [53].

RNN have shown great success in many NLP tasks. The most commonly used type of RNNs are long short term memory recurrent neural networks (LSTM RNN), which are much better at capturing long-term dependencies than typical recurrent neural network discussed before.

2.5.2. Long short-term memory neural networks (LSTM)

Sometimes it might be sufficient to remember recent information to perform recent task. But there are also cases where we need more context information. As the gap between relevant information and the point where it is needed becomes very large RNNs become unable to learn to connect the information. Long Short Term Memory networks usually just called 'LSTMs' are a special kind of RNN, capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the

long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn [54]. All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer. LSTM also shares this repeating structure, but contains four neural network layers interacting in a special way. The LSTM have the ability to remove or add information to the cell state, which is carefully regulated by structures called gates. Gates control the flow of information. They are composed of sigmoid neural network layer and appoint wise multiplication operation. The sigmoid layer outputs numbers between 0 and 1 which describes how much of each component should be let through LSTM's has three such gates to protect and control the cell state. The first step in LSTM is to decidewhat information to throw away from the cell state. This decision is made by a sigmoid layer called the forget gate layer. It looks at h_{t-1} and x_t , and outputs a number between 0 and 1 for each number in the cell state C_{t-1} . A 1 represents ‘completely keep this’ while a 0 represents ‘completely get rid of this’ [53]. LSTM decision process diagram on what information to throw is shown in Figure below.

$$f_t = \sigma(W_f \cdot [h_{t-1}; x_t] + b_f) \text{-----} 2.1$$

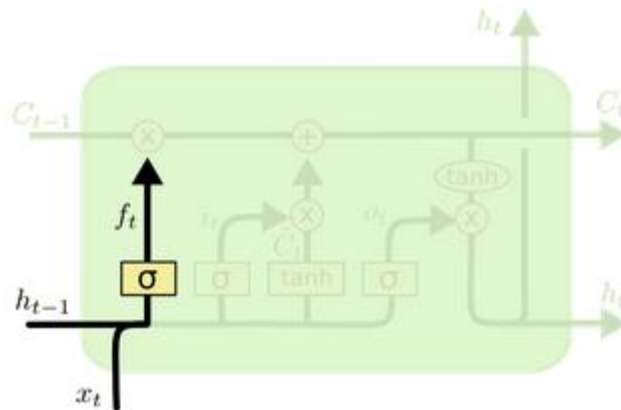


Figure 5. LSTM decision process on what information to throw

The next step is to decide what new information we’re going to store in the cell state. This has two parts. First, a sigmoid layer called the ‘input gate layer’ decides which values we’ll update. Next, a tanh layer creates a vector of new candidate values, $C_{\sim t}$, that could be added to the state. In the next step, we’ll combine these two to create an update to the state [53].

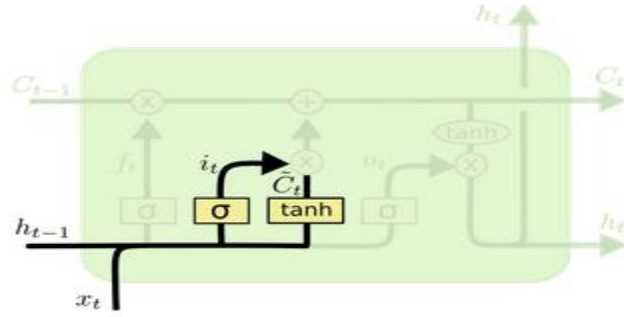


Figure 6. LSTM decision process on what information to store

$$i_t = \sigma(W_i \cdot [h_{t-1}; x_t] + b_i) \text{-----2.2}$$

$$C_t = \tanh(W_c \cdot [h_{t-1}; x_t] + b_c) \text{-----2.3}$$

Multiplying the old state by f_t , forgetting the things decided to forget earlier. Then adding $i_t * C_t$. This is the new candidate values, scaled by how much decided to update each state value. Finally the output will be based on the cell state. Sigmoid layer decides what part of cell state to output, then puts cell state through \tanh and multiply it by output of sigmoid gate.

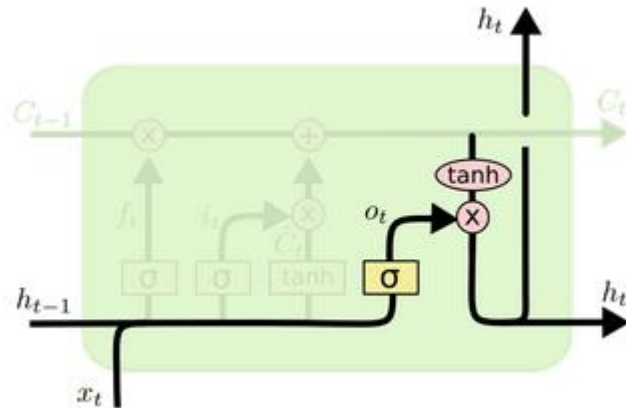


Figure 7. LSTM cell state output

$$O_t = \sigma(W_o [h_{t-1}; x_t] + b_o) \text{-----2.4}$$

$$h_t = O_t \cdot \tanh(C_t) \text{-----2.5}$$

2.5.3. Bidirectional Long short-term memory neural networks (BLSTM)

Bidirectional recurrent neural networks are extended form of regular long-short term memory neural networks [73]. Bidirectional long-short term memory (BLSTM) connect two hidden layers running in opposite directions to a single output, allowing them to receive information from both past and future states. This generative deep learning technique is more common in supervised learning approaches, rather than unsupervised or semi-supervised because how difficult it is to calculate a reliable probabilistic model.

2.5.4 Encoder-Decoder Sequence-to-Sequence Architectures

Deep Neural Networks (DNNs) are extremely powerful machine learning models that achieve excellent performance on difficult problems such as speech recognition [55]. Many important problems are best expressed with sequences; for example, speech recognition and machine translation. Likewise, question answering can also be seen as mapping a sequence of words representing the question to a sequence of words representing the answer. It is therefore clear that a domain-independent method that learns to map sequences to sequences would be useful. Sequences pose a challenge for DNNs because they require that the dimensionality of the inputs and outputs is known and fixed. Long Short-Term Memory (LSTM) architecture can solve general sequence to sequence problems. The idea is to use one LSTM to read the input sequence, one time step at a time, to obtain large fixed dimensional vector representation, and then to use another LSTM to extract the output sequence from that vector. The LSTM's ability to successfully learn on data with long range temporal dependencies makes it a natural choice for applications.

2.6 Summary

This chapter discussed overview of morphology and morphemes. Morphological segmentation and its approaches are discussed in detail. Additionally, neural networks, deep learning, and different kinds of deep learning are explained. The next chapter will discuss the related works for different natural languages.

CHAPTER THREE

3. RELATED WORKS

Morphological segmentation is initial step in different NLP applications and also one of the most popular research areas in the field of Natural Language Processing. Some of the research works have been done in different languages. Recently, most works were used neural network models. In this chapter we reviewed the previous works done on morphological segmentation.

3.1. System for Morphological Processing of Amharic, Afaan Oromo and Tigrinya (HornMorph)

The work of [15] on Afaan Oromo, Amharic and Tigrinya is another research presented on the area of morphology. A set of Finite State Transducers (FST) was applied in this research. FSTs were used for morphological analyzer and generator. The analyzer segment words into their morphemes and assign grammatical features. As an example, given Afaan Oromo words (noun or verb), hornmorph display the root, its first morpheme and feature description for each possible analysis. On the other hand, morphological generator performs reverse process; for given root of word, display meaningful word.

The author collected lexicon for three languages from online dictionaries. For Afaan Oromo, as the author discussed the lexicon verb and noun roots are extracted from the dictionaries of Gragg and Bitima that contains contain 4,112 verb roots and 10,659 nouns stems. For Tigrinya, the lexicon of verb roots is derived from Efrem Zacarias around 602 verb roots. Likewise, For Amharic, lexicon is derived from the Amharic-English dictionary of Aklilu which contains 1,851 verb roots and 6,471 noun stems.

The morphological analysis of [15] has two versions; anal_word that used for analyzing single word and anal_file used for analyzing multiple words. Both of them take input words and output a root or stem and grammatical analysis. For Amharic and Afaan Oromo there are two additional analysis functions; seg_word and seg_file which segment inputs (also nouns for Afaan Oromo into sequences of morphemes. For Amharic only, there are additional functions; phon_word and phon_file which convert orthographic form to a phonetic form as required for text-to-speech applications. Based on those functions author evaluated the HornMorph with 200 Amharic and

Tigrigna verbs, and 200 Amharic nouns and adjectives. Each word was selected randomly. The `anal_word` function was run on those words and the results were evaluated by a human reader who is familiar with the languages. The program made 8 (96% accuracy) and 2 (99% accuracy) errors for Tigrinya and Amharic verbs respectively. For Amharic nouns and adjectives it made 9 errors (95.5% accuracy). For Afaan Oromo language, the authors were not put the result of performance of the system.

3.2. Suffix Sequences Based Morphological Segmentation for Afaan Oromo

Getachew [12] has been developed a morphological segmentation for Afaan Oromo using machine learning approach. As the study explained, the author used suffix sequences for detecting morpheme boundaries of Afaan Oromo words. The author was proposed to utilize divisive hierarchical clustering and distribution frequency to build hierarchical candidate stem tree from which segmented suffix sequences can be modeled. Suffix sequences are series of suffixes attached to the root of a word. One factor that increases the number of word forms of a language is the complexity of its suffix sequences because one suffix sequence could be attached to different roots for forming different word forms.

In machine learning-based approaches, informative linguistic structures of a language are learned from corpora. Due to there is no compiled corpora for Afaan Oromo morphological analysis, the author collected dataset for training and testing word list. The compiled wordlists are based on the language morphological features. Finally, the author compiled 3370 unique words as training word list. For testing word list, author compiled 6000 word types, list of unique types from different sources mainly from internet. Among them 500 words were randomly selected to develop the testing word-list. Lastly, author trained, tested the model and accuracy achieved was 94.3% F-measure.

3.3. Morphological Segmentation with LSTM neural networks for Tigrigna language

Tigrinya belongs to the Semitic language branch of the Afro-asiatic family, along with Hebrew, Amharic, Maltese, Tigre, and Arabic. Tigrinya is a language spoken by over 7 million speakers in Eritrea and northern Ethiopia. The writing system, known as the Ge'ez script, is

adopted from the ancient Ge'ez language, which is currently used as a liturgical language. The Ge'ez script is an abugida system in which each letter (alphabet) represents a consonant-vowel (CV) syllable. The Tigrinya alphabet chart, known as 'Fidel', comprises of about 275 symbols.

Morphological segmentation using LSTM neural networks for Tigrigna were conducted by Yemane [18]. The study used Conditional random fields (CRF) and window-based long short-term memory (LSTM) neural networks were employed separately to develop boundary detection models. The author constructed a new morphologically segmented corpus with about 45,127 manually segmented tokens. The study demonstrated the problem is addressed by sequence tagging task and proposes a fixed-size window approach for modeling contextual information of characters. The author used Keras to develop and Hyperas to tune our deep neural networks. Using these resources the accuracy achieved was 94.67% F1 score using bidirectional LSTMs with window approach to morpheme boundary detection.

3.4. Morphological Analysis of Ge'ez Verbs Using Memory Based Learning

Ge'ez is the classical language of Ethiopia which belongs to the Semitic language. Still now, it is used as the liturgical language of EOTC. Ancient philosophy, history, tradition and knowledge were written by ge'ez. To automatically analyse these documents, ge'ez morphological analyzer should be developed.

Yitayal [60] proposed the model to automatically analyse these documents. Morphological analyzer is one basic tool in automatic processing of any human language. The author was used memory based learning to automatically analyses morphology of ge'ez verbs. The system has two major components: Training and Analysis. In training phase, annotation process was identified for dataset in a character based representation of the features. Then, these annotated dataset are extracted in a fixed length of instance vectors using windowing method. Next, instances are passed to the memory based learning tool (TiMBL). Finally, the learning model is built. The analysis phase performs instance making by extracting features from the given text to have similar structure of features during comparison. Then the extracted features are passed to the morpheme identification process to be compared with individual instances in memory and stems are extracted with their morpheme functions. Finally, the roots are extracted from the stems.

The model was implemented using python programming language and was used TiMBL's algorithms; IB2 and TRIBL2. 10-fold cross validation technique was used for evaluation of performance. The overall accuracy with optimized parameters using IB2 and TRIBL2 was 93.24% and 92.31%, respectively. Therefore, IB2 algorithm shows better result than TRIBL2 algorithm for Ge'ez verb morphology. The author was concluded, as the number of the training dataset increase, the accuracy of the unseen data can be increased.

3.5. Morphological Segmentation with Seq2Seq neural network for Russian language

The study was done by Arefyev [17] with the objective of developing morphological segmentation using English and Russian datasets. For English dataset authors were used the same training dataset with [80] and for Russian dataset the same training with [58] were used. The datasets were used for training and evaluating morphological segmentation algorithms. As the study demonstrated the method used was based on the sequence to sequence neural model. Sequence to sequence (seq2seq) is general-purpose neural network architecture for sequence transduction, which is used for tasks such as machine translation, text summarization, conversational modeling and more as described Denny Britz et al [56]. The author was adapted seq2seq, consisting of an encoder and decoder, with an attention mechanism for morphological segmentation. An encoder reads in source data such sequence of symbols, and produces vector containing information about this data relevant for the task. A decoder is a generative model that is conditioned on the representation created by the encoder. Instead of encoding the input sequence into a single fixed size representation the model can, with attention mechanism [57], learn how to generate an input representation for each output time step. In other words, the model learns which elements of the input sequence to attend to in order to generate the next element of output sequence, based on the input sequence and what it has produced so far.

For training the model, morphological segmentation task was defined as sequence transduction, that is, the sequence of symbols is being transformed into another sequence of symbols. For this purpose, every word in training datasets was represented as the sequence of its letters, for example in English (*windowless*). Additionally, the special symbol "*" was added into target training dataset. This symbol indicated the boundaries between word's segments (*window*less*). The study was achieved 0.9395 accuracy using seq2seq model.

3.6 Neural Morphological Analysis: Encoding-Decoding Canonical Segments

Katharina [59] proposed a character-based neural encoder-decoder model. The author extends the seq2seq model developed by including morpheme level and lexical information through a neural reranker. The study considered canonical segmentation where each word is divided into a sequence of standardized segments. But, most research has traditionally focused on surface segmentation, where by a word w is segmented into a sequence of substrings whose concatenation is the entire word [38]. To make the difference concrete, consider the following example: the surface segmentation of the complex English word *achievability* is *achiev+abil+ity*, whereas its canonical segmentation is *achieve+able+ity*, means the author developed model to restore the alterations made during word formation. The goal of the study was to map a word $w \in \Sigma^*$ (e.g., $w=achievability$), to a canonical segmentation $c \in \Omega^*$ (e.g., $c=achieve+able+ity$). We define $\Omega = \Sigma \cup \{+\}$, where $+$ is a distinguished separation symbol. Additionally, the author indicated the segmented form as $c=\sigma_1+\sigma_2+\dots+\sigma_n$, where each segment $\sigma_i \in \Sigma^*$ and n is the number of canonical segments.

The proposed model consists of two parts as [57]: First, an encoder-decoder recurrent neural network (RNN) to the sequence of characters of the input word to obtain candidate canonical segmentations. Second, author defined a neural reamer that allows us to embed individual morphemes and chooses the final answer from within a set of candidates generated by the encoder-decoder. The dataset used was prepared by [31] for canonical segmentation. It was consists of three languages namely, English, German and Indonesian. The author trained the model and achieved accuracy greater than or equal to 94 % (≥ 94). The following table indicate summarization of related works.

Author	Approach	Corpus size		Language	Performance (%)
		Training size	Testing size		
[12]	Suffix-sequences	3370	500	Afaan Oromo	94.3
[15]	FSTs	1851 verb roots and 6471 noun stems	200 verbs and 200 nouns and adjectives	Amharic	99% for verbs and 95.5% for nouns and adjectives
		602 verb roots	200 verbs	Tigrinya	96%
		4,112 verb roots and 10,659 nouns stems	--	Afaan Oromo	--
[18]	CRF and LSTM	45,127		Tigrinya	94.67
[60]	Memory based Learning	994.5	110.5	Ge'ez	IB2=93.24 TRIBL2=92.31
[17]	Seq2Seq LSTM	English dataset (Sami Virpioja et al., 2013)	686	English	0.8865
		73,639	24,547	Russian	0.9395
[59]	Se2seq RNN	8000	2000	English, German and Indonesian	>=94

Table 1. Summary of related works

3.7 Summary

This chapter discussed about the related works of morphological segmentation. The methodologies used and performance of different models achieved were discussed. To the best of our knowledge, there is no research work done for Afaan Oromo morphological segmentation using Neural Networks. The next chapter will discuss the overview of Afaan Oromo writing system and its morphology. From the reviewed works, we have observed that morphology segmentation were preprocessing step and recently morphological segmentation systems for various languages have been developed mostly using neural networks approach.

CHAPTER FOUR

4. OVERVIEW OF AFAAN OROMO

4.1 Introduction

The Oromo language (also known as Afaan Oromoo) is a Cushitic language spoken by about 40 million people in Ethiopia. It is also widely spoken in Ethiopia and some neighboring countries like Kenya and Somalia [8, 62]. It is one of the five most widely spoken languages among the thousands of languages spoken in Africa [63]. According to the census taken in 2007, the number of people speaking Afaan Oromo as their mother tongue is 34.5% of total population of Ethiopia [9]

Nowadays, Afaan Oromo is an official language of Oromia regional state of Ethiopia and also academic language for primary school of the region. Moreover, there are journal, magazines, newspapers, news, online education, books, entertainment Medias, videos, pictures, are available in electronic format both on the Internet and on offline sources are published in the language.

4.2 Afaan Oromo Writing System

The Afaan Oromo writing system is a modification to Latin writing system. The writing system of the language is known as “qubee Afaan Oromo” is straightforward which is designed based on the Latin script. The ‘*Qubee*’ writing system has a total of 33 letters of which 26 of them are similar with English letters and 7 of them are combined consonantal letters known as ‘*qubee dachaa*’ (digraphs). The digraphs include ‘*ch*’, ‘*dh*’, ‘*sh*’, ‘*ny*’, ‘*ts*’, ‘*ph*’ and ‘*zy*’ [64,65]. There are five vowels in ‘*Qubee Afaan Oromo*’. These are ‘*a*’, ‘*e*’, ‘*i*’, ‘*o*’, ‘*u*’. These vowels may appear as short vowels or long vowels in the language. A vowel is said to be short if it is one, For example, lafa /ground/. It is called a long vowel if it is two, which is the maximum, For example, lafaaa /soft/. The use of long and short vowels can result in different meanings.

There are 28 consonants in ‘*Qubee Afaan Oromo*’ including digraphs. Geminating (doubling a consonant) is also significant in Afaan Oromo because consonantal length can distinguish words from one another as the following examples. Most Afaan Oromo consonants can be geminated except ‘*h*’ and digraphs [66]

Word	Meaning	Geminated word	Meaning
Guura	Gather	Gurra	Ear
Baala	Leaf	Ballaa	Blind
Boowa	Cry	Bowwaa	Cliff
Dame	Branch	Dammee	Sweety

Table 2 Word and its geminated form in Afaan Oromo

The digraphs are count as single consonants though they are written as two letters. Afaan Oromo vowels and consonants with their sound are shown in Appendix-2.

4.3 Afaan Oromo Morphology

Morphology is the study of morphemes and their arrangements in forming words. Morphemes are the minimal meaningful units which may constitute words or parts of words [68]. Like a number of other African and Ethiopian languages, Afaan Oromo has a very rich morphology. It has the basic features of agglutinative languages where all bound forms (morphemes) are affixes. In agglutinative languages like Afaan Oromo, Amharic and Zulu most

of the grammatical information is conveyed through affixes (prefixes, infixes and suffixes) attached to the roots or stems.

In Afaan Oromo, words can be formed from morphemes in two ways [66]: inflectional and derivational. Inflectional morphology is concerned with the inflectional changes in words where word stems are combined with grammatical markers for things like person, gender, number, tense, case and mode. Inflectional changes do not result in changes of parts of speech. Derivational morphology deals with those changes that result in changing classes of words (changes in the part of speech). For instance, noun or an adjective may be derived from a verb.

4.3.1 Types of Morphemes in Afaan Oromo

There are two kinds of morpheme; *free morpheme and bound morpheme*. Free morpheme is a simple word consists of single morpheme and a morpheme with the potential for independent

occurrence. Bound morpheme requires another morpheme to make up a word; they can't occur independently. In Afaan Oromo roots like 'muk-'/tree/,' laf-'/ground/ is bound morpheme as they are parts of the word. Like the root, an affix is also a morpheme that cannot occur independently. It is attached in some manner to the root, which serves as a base. These affixes are of three types- prefix, suffix and infix. Affixes can take the form of a prefix at the beginning of a root word, or as a suffix at the end of a root word. In fageenya /distance/, -eenya is a suffix and fag- is a root of word. An *infix* is a word element (a type of affix) that can be inserted within the base form of a word rather than at its beginning or end to create a new word. But, Afaan Oromo does not have infixes [66].

There are a wide range of word formation processes in Afaan Oromo. Since Afaan Oromo is morphologically productive almost all nouns, pronouns and determinants, case and relational concepts, functional words, verb and adverbs can be affixed.

4.3.1.1 Noun Morphology

There are three major types of suffixes in Afaan Oromo: Derivational, inflectional and attached suffixes. Afaan Oromo attached suffixes are particles or postpositions like -arra, -bira, -irra, -itti and -dha while inflectional suffixes comprises the most frequent and dominant suffixes such as – n, -lee, -een, -icha, -tu, -oo, -oota and -wwan.

In Afaan Oromo derivational suffixes are –ummaa, -ina, -eenya and –achuu...etc. Derivational suffixes are often used for formation new words in the language following stem or base form of other words. In complex word structure, certain set of suffixes conventionally come in sequence in which one suffix come before another suffixes. The most common order/sequence of the above major three Afaan Oromo suffixes (within a given word) is: <stem><derivational suffixes><inflectional suffixes><attached suffixes> [69].

I. Noun Inflections

Afaan Oromo nouns are words used to name any of categories of things, people, places or ideas. Almost all nouns in Oromo end with a vowel except for a few of them which end in specific consonants like *n*, *l*, *t*. Nouns are inflected to indicate different grammatical functions such as

number, gender, definiteness and case. The principles of noun inflection here apply to nouns and adjectives.

a) Pluralization

Pluralization indicates quantity of things in numbers. If something is one, we can say single, in Afaan Oromo /qeenxee/ and if it is more than one we can say many(/danuu/ in Afaan Oromo). But, all of the nouns cannot be marked by this quantity. A singular is marked by zero morphemes where as a plural noun is marked morphologically by suffixing the morpheme like -oota, -oolii, -een, -lee, -wwan, -yyii, -eetii, -ii, -oo to the base as free alternates [69].

Nouns which can be suffixed to form plural are countable nouns such as nama/person/, biyya/country/, muka/tree/, harree/donkey/, mucaa/baby/, hojii/work/, gaara/mountain/, laga/river/, quba/fingers/...etc. Nouns which cannot be suffixed to form plural are called uncountable nouns; such as bishaan/water/, daakuu/.../, ibidda/fire/...etc. Sometimes there are countable nouns which cannot be pluralize, such as ija/eye/, luka/foot/, ilkaan/tooth/, gurra/ear/, harka/hand/. Those nouns are pluralized as ija lamaan /one eye/, luka tokko/ one foot/ because these nouns especially nouns of human body or animal body can be accepted as the part of the body.

Generally, there are nine commonly used pluralization affixes in Afaan Oromo; -oota, -oolii, -een, -lee, -wwan, -yyii, -eetii, -ii, -oo

The first category of suffixes, -oota, -oolii, -oolee, is affixed to virtually any noun to form plural. But all the terminal vowels, except nouns having the phoneme /o/ as the last vowel(s), in the citation form get deleted when these plural formative suffixes are attached to nouns. These suffixes are used with nouns that end with both long and short vowels. The following examples correspond to the three suffixes.

Singular nouns	Gloss	Plural nouns	Gloss
Barataa	Student	Barattoota	Students
Keessummaa	Guest	Keessummoota	Guests
Farda	Horse	Fardoolii	Horses
Warra	Family	Warroorii	Families
Naannoo	Region	Naannoolee	Regions

Table 3. First category of pluralization suffixes

The second category of suffixes to pluralize nouns are –lee and –wwan. Nouns which can pluralize with suffixes –lee and –wwan do not delete final vowels and they are used with nouns terminated with long vowels as the following examples.

Singular nouns	Gloss	Plural Nouns	Gloss
Gaaffii	Question	Gaaffiiwwan	Questions
Garee	Group	Gareewwan	Groups
Urjii	Star	Urjiilee	Stars
Warshaa	Factory	Warshaalee	Factories

In Afaan Oromo nouns pluralized by the suffixes set -oota, -oolii, -oolee in long form or -ota, -olii, -olee in short form depend on the vowel present in the syllable that precedes the last syllable. If the vowel in the syllable that precedes the last syllable is short, the noun takes -oota as in the case of fira(relative) becomes firoota (relatives), while the noun having long vowel in the syllable prior to the last syllable takes the suffix -ota as in the case of barsiisaa (teacher) becomes barsiisota (teachers).

The third category of suffixes that behave differently from the discussed above are –an and –een. All nouns that can take suffix –an and -een to make plural noun by doubling consonant of the last syllable. These nouns mostly end in the consonantal phoneme l, m and r in the case of –an, and g, k, n in the case of -een followed by short vowel. Following are some examples.

Singular nouns	Gloss	Plural Nouns	Gloss
Daa'ima	Baby	Daa'imman	Babies
Wasiila	Uncle	Wasiillan	Uncles
Beera	Old woman	Beerran	Old women
Gaala	Camel	Gaallan	Camels
Laga	River	Laggeen	Rivers
Muka	Tree	Mukkeen	Trees

The fourth category of suffixes to pluralize nouns are –eyyii.. The suffix –eyyii forms plural by deleting complex endings that form nouns like –eessa, eensa, eettii and attaching the plural maker -eyyii.

Singular nouns	Gloss	Plural Nouns	Gloss
Hiyyeessa	Poor	Hiyyeeyyii	Poor
Bineensa	Animal	Bineeyyii	Animal
Sooressa	Rich	Sooreyyii	Rich
Waraabessa	Hyena	Waraabeyyii	Hyenas

As shown in the above examples, if the vowels found in the noun stem is long such as Waraab, Soor the plural maker would be –eyyii and if the vowels in noun stem is short it takes –eeyyii plural maker.

The fifth category of suffix to pluralize nouns is –oo . The suffix –oo forms plural by changing stem of the word. Following are some examples.

Singular nouns	Gloss	Plural Nouns	Gloss
Farda	Horse	Faradoo	Horses
Dubra	Girl	Dubaroo	Girls

The sixth category of suffix to pluralize nouns –eetii. Afaan Oromo nouns can be pluralized by –eetii if vowel found on last word is different from beginning vowel of suffix. Sometimes they double last consonant of the word as suffix –een

Singular nouns	Gloss	Plural Nouns	Gloss
Mana	Home	Manneetii	Homes\\
Muka	Tree	Mukkeetii	Trees

The seventh category of suffix to pluralize nouns is –ii. There are several words which take this plural maker. The nouns which can take suffix –ii can be:

1. As the nouns which can take suffix –oo they can change stem of the word and
2. As the nouns which pluralized by –een they can double the consonant of the word.

Singular nouns	Gloss	Plural Nouns	Gloss
Korma	Bull	Korommii	Bulls
Goromsa	Cow	Gorommii	Cows

Generally, there may be additional suffixes exist in Afaan Oromo for different dialects. All of the plural makers discussed above cannot be applied for all nouns, but a noun can be pluralized with different suffixes. For example, saroota, sareewwan, saroolii, sarootii, means in English dogs. Hence a noun can be pluralized with two or more than two different suffixes as given example. It is difficult to categorize nouns according to the suffixes they take for making plural as the way plural makers are attached to the nouns may or may not be similar.

Category no.	Suffix	Its description
1	-oota,- oolii, oolee	Pluralize noun by deleting the last vowel
2	-wwan, -lee	Pluralize noun without deleting the last vowel
3	-een, -(a)n	Pluralize noun by doubling last consonant
4	-eeyyii	Pluralize noun by dropping –eessa/eensa
5	-oo	Pluralize noun by changing stem of the word.
6	-eetii	nouns can be pluralized by –eetii if vowel found on last word is different from beginning vowel of suffix.
7	-ii	Nouns can be pluralized by –ii can be: <ol style="list-style-type: none"> 1. As the nouns which can take suffix –oo they can change stem of the word and 2. As the nouns which prularized by –een they can double the consonant of the word.

Table 4. Afaan Oromo pluralization suffixes summarization

b) Singulative

Singulative is one of the suffix that applied to the noun to provide different services. The services that provided by singulative is to differentiate object/person from other similar object/person. Additionally, it adds more information to the object/person. The singulative marker shows that noun is marked for being used as single form which may or may not be definite. In Afaan Oromo there is no indefinite articles (*a* and *an* in English), but it indicates definiteness (English the) with suffixes on the noun: *-icha* for masculine nouns and *- ittii* for feminine nouns. Vowel endings of nouns are dropped before adding these suffixes

Baseform	Inflected Form	Meaning
Mana	Manicha	‘the house’
Jaarsa	Jaarsicha	‘the old man’
Jaartii	Jaartittii	‘the old woman’
Dubartii	Dubartittii	‘the woman’

Table 5. Afaan Oromo suffixes for singulative indicator

Both the morphs *-icha* and *-ittii* tend to be singulative markers embodying the property of definiteness.

c) Gender

Like most other Afro-asiatic languages, Oromo has two grammatical genders, masculine and feminine, and all nouns belong to either one or the other [70]. These are identified through gender marking suffixes, or lexically by using different words for masculine and feminine forms. The distinct words for masculine and feminine like *adaadaa* „aunt“ and *eessuma* „uncle“ are also used in Oromo. Gender indicating words can be used for animals and they are placed immediately after or before the nouns they belong to. The most common contrastive pair of words used in this way is *kormaa*, *male* (m.)“ Vs. *dhaltuu* , *female* (f.). Consider the table below.

Table Gender Distinction in Nouns

Base form	Masculine		Feminine	
	<i>Inflected form</i>	<i>Meaning</i>	<i>Inflected form</i>	<i>Meaning</i>
Gurbaa	Gurbaa	‘Boy’	Intala	‘Girl’
Abbaa	Abbaa	‘Father’	Haadha	‘Mother’
Cimaa	Cim-aa	‘clever boy’	Cimtuu	‘clever girl’
Jabaa	Jab-aa	‘strong boy’	Jabduu	‘strong girl’
Sooressa	Soore-ssa	‘rich man’	Soore-ttii	‘rich girl/woman’
ogeessa	ogee-ssa	‘technician’	ogee-ttii	‘technician lady’

Table 6. Afaan Oromo suffixes for gender indicator

In above Table, the first two examples are distinguished for gender lexically. The third and fourth nouns that are derived from verbs indicate that the long *-aa* suffixed to the verb root to marks masculine gender whereas the suffix *-tuu/-duu* makes verbal nouns in feminine gender.

Proper nouns may also lexically gender coding nouns distinguish between masculine and feminine genders by their contrastive final syllables as *-ssa* vs. *-tii*. Such nouns that are derived from adjectives indicating gender distinction. Proper nouns may also code gender distinction by varying their final vowel like *Lalisaa* (m.) vs. *Lalisee* (f.) in Oromo.

d) Cases

The relational category, case, is a grammatical relationship of nouns or pronouns to other words in a sentence. Case is a grammatical category of nouns that indicates the nature of their relationship to the verb in sentences [71, 23].The number of cases varies from language to language. In this regard, nouns in Afaan Oromo are inflected for nominative, ablative, instrumental and locative cases.

i. Nominative case

The nominative is used for nouns that are the subjects of clauses. The nominative case is marked by four different morphs of allomorphic variation occurring in complementary distribution. The allomorphs for the nominative case are *-n*, *-ni*, *-i* and \emptyset .

The difference in the phonological realization of the nominative case markers arises from the phonological nature of the nouns. The marker *-n* occurs after a terminating long vowel of a noun including the derived nouns. If a noun base that ends in short vowel, the final vowel is dropped and *-ni* is suffixed to mark nominative case. The allomorph *-i* is appended to two consonants or a geminated consonant. Nominative case can also be marked by zero morph when the noun ends in consonant as in the last noun in the following table.

Base forms	Inflected Forms	Meaning
Lafa	laf-ni	‘land’
Maqaa	Maqaa-n	‘name’
Ibsa	Ibsi	‘description’
Morma	Morm-i	‘neck’
Afaan	Afaan-0	‘Language’

Table 7. Afaan Oromo suffixes for nominative indicator

The marker **-ti** is used as a phonological variant of the morpheme **-ni** in nominative case. The nominative case allomorph **-ni** undergoes phonological processes and gets changed to **-ti**, for example **'bofa'**snake - **bof-ni** snake Nom- **bof-ti** snake Nom in which **-ti** results from partial assimilation process in the shares vocal feature between the segments /f/ and /t/.

The marker **-tu** is contrastive focus marker for nominal because it contrasts the focused nominal with other presupposed constituents providing context.

For example, *gurbichatu na waame.* /the boy who called me/

ii. Genitive

The genitive is used for possession. It corresponds roughly to English *of* or *'s*. The genitive is usually formed by lengthening a final short vowel, by adding **-ii** to a final consonant, and by leaving a final long vowel unchanged. The possessor noun follows the possessed noun in a genitive phrase.

Examples

- ✓ *obboleetti* 'sister', *namicha* 'the man', *obboleetti namichaa* 'the man's sister'
- ✓ *barumsa* 'field of study', *afaan* 'mouth, language', *barumsa afaanii* 'linguistics'

In place of the genitive it is also possible to use the relative marker *kan* (m.) / *tan* (f.) preceding the possessor. For Example, *obboleetti kan namicha* 'the man's sister'

iii. Dative

The dative is used for nouns that represent the recipient (*to*) or the benefactor (*for*) of an event. The dative form of a verb infinitive (which acts like a noun in Oromo) indicates purpose. The dative takes one of the following forms.

1. Lengthening of a final short vowel (ambiguously also signifying the genitive)
For example *manicha* 'the house', *manichaa* 'to the house, for the house'
2. *-f* following a long vowel or a lengthened short vowel; *-iif* following a consonant as the following examples
 - ✓ *Lafa* /land/, *lafaaf* /to a land/

- ✓ *Mana* /house/, *Manaaf* /to a house/
- ✓ *Deemuu* /to go/, *Deemuuf* /in order to go/
- ✓ *bishaan* /water/, *bishaaniif* /for water/

3. dhaa or -dhaaf following a long vowel

For example saree 'dog'; sareedhaa, sareedhaaf 'to a dog'

4. -tti (with no change to a preceding vowel), especially with verbs of speaking

For example Caalaa man's name, himi 'tell', Caalaatti himi /tell Caalaa/

iv. Instrumental

Using of instruments or a means of doing something is termed as the instrumental case. It is used for nouns that represent the instrument ("with"), the means ("by"), the agent ("by"), the reason, or the time of an event.

The suffixes -n, tiin following long vowel or lengthened short vowel, -iin following a consonant, and -dhaan following a long vowel indicates instrumentation as following examples in the table.

Base forms	Meaning	Inflected forms	Meaning
Harka	'Hand'	Harkaan	'by hand'
Nalkan	'Night'	Halkaniin	'at night'
Afaan Oromo	'Afaan Oromo'	Afaan Oromootiin	'in Afaan Oromo'
Yeroo	'Time'	Yeroodhaan	'on time'
Bawuu	'To come out'	Bawuudhaan	'by coming out'

Table 8. Afaan Oromo suffixes for instrumental indicator

v. Locative

The locative is used for nouns that represent general locations of events or states. It is marked by the suffix -tti, and tells location for some occurrence, goal or address.

Examples:

Base forms	Meaning	Inflected forms	Meaning
Mana	'Home'	Manatti	'At home'
Magaalaa	'Town'	Magaalaatti	'in Town'
Aangoo	'Authority'	Aangootti	'by Authority'
Harka	'Hand'	Harkatti'	'in Hand'

Table 9. Afaan Oromo suffixes for locative indicator

vi. Ablative

The ablative is used to represent the source of an event; it corresponds closely to English *from*. The ablative, applied to postpositions and locative adverbs as well as proper nouns, is formed in the following ways:

- ✓ Vowel length is about lengthening of a short vowel, especially, referring to long *-aa* and *-ii* to mark ablative case.

For example **keessa** 'inside, in', **keessaa** 'from inside'

Jimma 'Jimma', **Jimmaa** 'from Jimma'

Shanan 'shanan' **Shananii** 'from shanan'

- ✓ Nouns that end in long vowels are marked for ablative case by placing copulas *-dhaa*.

For Example **Adaamaa** 'Adama', **Adaamaadhaa**, 'from Adama'

Magaalaa 'City', **Magaalaadhaa**, 'from City'

- ✓ When the word ends in a consonant, *-ii* is added
For Example, **Harar** 'Harar' **Hararii** 'from Harar'

- ✓ Following a noun in the genitive, *-tii* is added.
mana 'house', **buna** 'coffee', **mana bunaa** 'cafe', **mana bunaatii** 'from café'

An alternative to the ablative is the postposition **irraa** 'from' whose initial vowel may be dropped in the process:

For example, **biyya** 'country', **biyya irraa**, **biyyarraa** 'from country'

vii. Vocative

Some languages have vocative case which marks the noun representing the entity (animate) we address. It is a verbal means of calling attention.

Base Form	Inflected form	Meaning
nama	nama-na	‘(you) guy’
Jara	Jara-na	‘(you) guys’
Gurbaa	Gurbaa-nana	‘you boy’
Bara	Bara-na	‘this year’

Table 10. Afaan Oromo suffixes for vocative indicator

Based on the above table:

- ✓ The suffix *-na* which marks vocative case is appended to a noun which is two syllabic and ending in short vowel with harmonic occurrence of vowels.
- ✓ Its full word form *nana* is used after nouns that end in long vowel.
- ✓ Sometimes the suffix *-na* can be used representing the word *kana* ‘this’

4.3.1.2 Noun Derivation

In Afaan Oromo as there are nouns without derived from another class exist, there also nouns derived from different word classes. Derived nouns are produced by adding suffixes to different word classes. Hence, the Oromo language is very productive in word formation by different means. One of the methods is the use of different derivational suffixes. In Afaan Oromo, derivational suffixes enable a new word, often with a different grammatical category to be built from stem/root of other words. A new form of nouns can be created by adding suffixes to the other nouns, adjectives, object..etc There are three process of driving a noun from other word classes. The following sections discuss processes in detail.

1) Nouns Derived from another nouns

New nouns can be derived from other nouns by adding the suffixes *-ummaa*, *eenya*, *ooma* to the noun stems. Thus, when these abstract noun formative morphemes are added to nouns, the final vowels of these words are deleted as the following set of examples illustrate.

Noun	Gloss	derived noun	Gloss
Nagaa	Peace	Nageenya	Peaceful
Guddaa	Great	Guddummaa	Greatness
Waaqa	God	Waaqummaa	Goddess
Fira	Relative	Firooma	Relativity
Olla	Neighbor	ollooma	Neighborhood

As it shown in the above table existing nouns and derived nouns are both found in one word class that is noun. Also when we see the meaning of them is very close to each other.

2) *Nouns Derived from Verbs*

As shown in the above table as suffix –eenya can be applied to the noun to derive abstract noun, it can also be applied to the verb to drive a new nouns as the following examples.

Verb	Gloss	Noun	Gloss
Qabaachuu	to have	Qabeenya	Property
Jiraachuu	to live	Jireenya	Life
Jabaachuu	to be strong	Jabeenya	Strength

Besides the suffix –eenya there is also another suffixes to drive nominal from verbs such as -aa,, -tuu, -ina, -noo, -ii, -ee, -iinsa, -iisa, -umsa, -maata, -aatii. The following examples illustrate them.

Verb	Gloss	derived noun	Gloss
Rakkachuu	to suffer	Rakkina	Problem
Hir'achuu	to decrease	Hir'ina	deficiency
To'achuu	to control	to'annoo	Control
Hubachuu	to understand	Hubannoo	Understanding
Barachuu	to learn	barumsa	Education
Falmuu	to argue	Falmii	Argument
Tiksuu	to protect	Tiksee	Protection
Dalaguu	to work	Dalagaa	Work
Hoogganuu	to lead	Hoogganaa	Leader
Bulchuu	to manage	Bulchiinsa	Manager
Qotuu	to farm	Qotiisa	Farming
Furuu	to solve	Furmaata	Solution
Loluu	to fight	Loltuu	soldier

3) Nouns Derived from Adjectives

In Afaan Oromo the noun can be derived from adjectives by suffixing the morphemes like –ooma, -ina, -ummaa eenya, The following examples indicate the derivation of such nouns.

Addeessa	Gloss	Stem	Suffix	Derived Noun	Gloss
Fagoo	far	fag	-eenya	fageenya	distance
Goota	hero	goot-	-ummaa	gootummaa	heroic
Furdaa	fat	furd-	-ina	furdina	fattiness
Jabaa	strong	jab-	-ina	jabina	strength
Arjaa	wise	Arj-	-ooma	Arjooma	kindness

4.3.1.3 Pronoun Morphology

A pronoun can be used to replace noun or noun phrase. Pronouns are inflected for properties of number, gender, singulative and case like the noun inflection. They are categorized in to six categories. Namely: personal, demonstrative, possessive, reflexive, reciprocal, and interrogative pronouns. The following sections discuss some of them.

A. Personal pronoun

A **personal pronoun** is a short word we use as a simple substitute for the proper name of a person.

Personal pronouns are inflected for types of cases [11] based on the following table.

English	Base	Subject	Dative	Instrumental	Locative	Ablative	Possessive adjectives
I	ana, na	ani, an	naa, naaf, natti	naan	natti	narraa	koo, kiyya [too, tiyya (f.)]
you (sg.)	si	ati	sii, siif, sitti	siin	sitti	sirraa	kee [tee (f.)]
he	isa	inni	isaa, isaa(tii)f, isatti	isaatii n	isatti	isarraa	(i)saa
she	isii, ishii, isee, ishee	isiin, etc.	ishii, ishiif, ishiitti, etc.	ishiin, etc.	ishiitti, etc.	ishiirraa, etc.	(i)sii, (i)shii
we	nu	nuti, nu'i, nuy, nu	nuu, nuuf, nutti	nuun	nutti	nurraa	keenna, keenya [teenna, teenya (f.)]
you (pl.)	isin	isini	isinii, isiniif, isinitti	isiniin	isinitti	isinirraa	keessan(i) [teessan(i) (f.)]
they	isaan	isaani	isaanii, isaaniif, isaanitti	isaanii tiin	isaanitti	isaanirraa	(i)saani

Table 11. Personal pronoun

B. Reflexive Pronoun

A reflexive pronoun indicates the person who realizes the verb action is same with receipt of the action. They are formed by prefixing the particle *of(i)-* on possessive adjective forms of pronouns. Reflexive pronouns are inflected for several cases like nominative case as in *ofiisaa ijaare* “himself built it” and dative case as in *ofiikee-f* “yourself-Dat”. Dative case is marked by *-f* on the reflexive pronouns. The following are list of examples.

Base form	Reflexive form	Meaning
ana	ofiikoo	“myself”
si	ofiikee	“yourself”
isa	ofiisaa	“himself”
ishee	ofiishee	“herself”
nu	ofiikeeñña	“ourselves”
isin	ofiikeessan	“yourselves”
isaan	ofiisaanii	„themselves“

Table 12. Reflective pronoun

C. Demonstrative Pronouns

A demonstrative pronoun is a pronoun that is used to point something specific with in a sentence.

The inflection of the demonstrative pronoun marked for nominative case as in *sun-i mana* ‘that is a house’.

Base Form	Nominative Form	Meaning
kana	kun(i)	this:Nom “this”
sana	sun(i)	that:Nom “that”
Tana	Tun(i)	This: Nom “this”

Table 13. Demonstrative pronoun

D. Interrogative Pronouns

Interrogative pronoun is used to make asking question. They are inflected for the following several cases.

Case	Base form	Inflected form	Meaning
Dative	eenyu	eenyuu-f	‘for whom’
	maal(i)	maalii-f	‘for what’
Locative	Eessa	Eessatti	‘where’
	Eenyu	Eenyutti	‘to whom’
Genitive	eenyu	eenyuu	‘whose’
	maal	maalii	‘of what’

Table 14. Interrogative pronoun

4.3.1.4 Adjectives Morphology

An adjective is a word that describes or modifies a noun or pronoun. It specifies to what extent a thing is as distinct from something else [72]. The inflectional categories or properties of adjectives are the same with that of nouns. Adjectives are inflected for number, gender, singulative and case like nouns; however, sometimes they are marked differently from nouns. For instance, adjectives,

unlike nouns, are inflected by reduplication to mark plurality. When adjectives occur with nouns in sentences, number is marked on both of them. Nouns are marked for plurality, but adjectives are marked for number by reduplication of its initial syllable (CV, CVC), or by the plural suffix *-(o)ota*. Here different ways of marking plural adjectives with examples:

Form of inflection	Singular	Plural
Lexical coding	hiyyeessa ‘poor:m’ bayeessa ‘beautiful’ : m Qabeettii ‘rich:f.’	hiyyeeyyii ‘poor’ bayeeyyii ‘beautiful’ qabeeyyii ‘riches’
Reduplication	xiqqaa ‘small’ ‘dheeraa’ ‘tall’	Xixiqqaa, ‘smalls’ dhedheeroo ‘talls’
-(o)ota	dadhabaa ‘lazy’ cimaa ‘clever’	Dadhaboota ‘lazies’ cimoota ‘clevers’

Table 15. Plural Forms of Adjectives

In Afaan Oromo adjectives the masculine form terminates in one of the following suffixes *-aa*, *-eessa*, or *-(a)acha*, and the feminine form terminates in one of the following suffixes *-oo*, *-tuu*, *-eettii*, or *-aattii*. The inflectional categories or properties of adjectives are the same with that of nouns. Markers of singulative property *-icha* (for masculine) and *-ittii* (for feminine) occur on adjectives also as for nouns. For example, ”mukti /mukti/ dheerichi” /the long stick/. But, Singulative markers are not used on both a noun and adjective at the same time.

If a noun is marked for nominative case, an adjective following it will also be marked for the same case. The following examples clarify the point.

Inflected forms	meaning
Mukti dheeraan	a long stick
Namootni dhedheeroon	a long men

Case	Marker	Example
Singulative	-cha, -ichi	adicha ‘the white’ dheerichi ‘the tall’
Beneficiary	-f	Muraasaa-f ‘for few’ Guddaaf ‘for big/respected’
Absolutive	No marker	Guddaa ‘a big’ Gowwaa ‘a foolish’
Nominative	-n, -ni, -i	Muk-ni ‘a stick’ Gaarii-n ‘a kind’
Genitive	Vowel length	Hamaa ‘bad’ Gowwaa ‘a foolish’

Table 16. Case Inflection Realized on Adjectives

4.3.1.5 Verb Morphology

Verbs are words or compound of words that expresses action, a state of being and/or relationship between two things. Verbs are morphologically the most complex POS in Afaan Oromo, with many inflectional forms; numerous words with other POS are derived primarily from verbs.

There are two major criteria to identify verbs from other word categories: syntax and morphology. In the former case, verbs function as predicates in a simple sentence and they are found at the end of a sentence. In the latter case, the agreement of verb with the number, gender and/or person of the subject, proper case markers for the different nominal forms and expression of the tense, aspect of the verb and number, specificity of the of nouns are some of the important morphological constraints governing correct generation.

1) Verb derivation

An Afaan Oromo verb stem can be the basis for four derived voices, passive, causative, and reflexive, each formed with addition of a suffix to the stem [65, 69, 72].

I. Causative

It is formed by adding -s, -sis, or -siis to the verb stem, except that stems ending in -l add -ch. Verbs whose stems end in' drop this consonant and may lengthen the preceding vowel before adding -s. Stems ending with **dha** cannot be causativised by –s, -sis, -siis suffixes. Verbs whose stems end in' drop this consonant and may lengthen the preceding vowel before adding -s. For instance, beek- 'know', beeksis- 'cause to know, inform', beeksifne 'we informed'; ka'- 'go up, get

up', kaas- 'pick up', kaasi 'pick up (sing.)!'; gal- 'enter', galch- 'put in', galchiti 'she puts in'; bar- 'learn', barsiis- 'teach', nan barsiisa 'I teach'.

II. Passive

The Afaan Oromo passive corresponds closely to the English passive in function. It is formed by adding -am to the verb root. For instance, jijjiir- ‘change’, jijjiiram- ‘to be change’.

III. Autobenefactive

Autobenefactive is formed by adding a morpheme of reflexive, middle voice and hence autobenefactive which is appended to verb root. The Afaan Oromo autobenefactive voice of a verb V corresponds roughly to English expressions such as 'V for oneself' or 'V on one's own', though the precise meaning may be somewhat unpredictable for many verbs. It is formed by adding **-adh** and **-at** to the verb stem.

For example, gurgura ‘sell’, gurguradha ‘sell for oneself’, gurgurate ‘he sell something for himself’.

2) Verb Inflection

Different inherent and agreement grammatical categories account for the inflection of verbs in Oromo. The inherent ones are aspect, mood, and voice whereas the agreement properties include person, number, gender and case. The three main functional domains of inherent verb inflection in the Oromo language are aspect, mood, and voice with some indications of tenses [81].

A. Aspect

In Afaan Oromo past and non-past can be identified by aspectual property. The markers of perfective and imperfective aspects *-e* and *-a* respectively occur on main verbs. When tense is considered, the perfect tenses and progressive tenses occur with auxiliaries *jir-* (present form) and *tur-* (past form) with their forms of agreement in inflection. Hence, the agreement markers are suffixed to the auxiliary verbs. The auxiliary verb always follows the main verb in a sentence. The aspect is, mainly, noticed on the converbal or progressive verb forms. The following table is overview of perfective and imperfective aspects

Verb type	Root	Perf form	Gloss	Impf form
Action	kuf-	kuf-e	‘fell-Perf’	kuf-a
Auxiliary	ĵir tur-	ĵira (present) ture (past)	‘exist-Perf’ ‘exist-Perf’	ĵira (present) ture (past)

Table 17. Aspectual Distinction on Verbs

b. Mood

Mood is the attitude of the speaker towards an utterance. It is originally from the word „mode“ which means a specific way of doing something. Modality, which is also originated from ‘mode’, is more ideal and is about the existence of a particular way of speaking. In connection with the styles of speech which arises from involvement of feeling, Oromo has several types of moods such as **indicative**, **imperative**, and **jussive** moods.

Indicative mood involves making statements and asking questions constitutes the most common clause type in Oromo. For example, Boruun Kitaaba Bareesse. /Boru wrote a book/

Imperative Mood: In Oromo, the imperative begins by the object as it precedes the verb in word order of the language. Intransitive verbs are used at the beginning of the sentence in the form of the subject ‘you’ understood. For example, Kitaaba sana fidi. /Bring that Book/ Negative forms of verbs in imperative sentences occur in a little bit special way. The particle *hin* and the dependent suffix *-n* both mark negativity. The suffix *-aa* marks mood and plural number the verb of an imperative form. The following table show affirmative and negative verbs in imperative form.

Verb types	Affirmative Imperative							
	Person	Verb Root	Imp	Imperative verb	Gloss			
Action	2singular	Bit-	-i	Bit-i	(you sg) buy.			
	2plural	Bit-	-aa	Bit-aa	(you pl) buy.			
Autobeneficative	2singular	Bit-at-(Bit-adh-)	-u	Bit-at-u(Bit-dh-a)	(you sg) buy:ABen			
	2plural	Bit-at-(Bit-adh-)	-aa	Bit-at-uu(Bit-dh-aa)	(you pl) buy:ABen			
Verb types	Negative Imperative							
	Person	Neg.	Verb Root	Imp	Ne g.	Imp	Imperative verb	Gloss
	2sg	hin	bit-	-i-	-n	-	hin bit-i-n	'Don't buy'
	2pl	hin	bit-	-	-n	-aa	hin bit-i-n-aa	'Don't buy'
	2sg	hin	bit-at-(bit-at-)	-i-	-n	-	hin bit-at-i-n	'Don't buy:ABen'
	2pl	hin	bit-at-(bit-at)	-	-n	-aa	hin bit-at-i-n-aa	'Don't buy:ABen'

Table 18. Affirmative and Negative Verbs in the Imperative mood

Jussive mood is marked by the pre-verbal particle *haa* and the dependent suffix *-u* or *-i* on the verb. For example, “let the boy come”. Negative forms of the jussive sentences is similar across the pronouns used which means the verb in the negative jussive doesn't occur agreeing with subject in number and person. The negative jussive sentences are formed by the proclitic *hin* along with its coexisting dependent suffix *-n* on the verb.

Jussive Mood						
Person	Jussive	Root	Agr.	Asp.	Inflected form	Gloss
3sgm	haa	dhuf-	-∅-	-u	haa dhuf-u	'Let him come'
3sgf	haa	dhuf-	-t-	-u	haa dhuf-t-u	'Let her come'
3pl	haa	dhuf-	-an	-i	haa dhuf-an-i	'Let them come'
1pl	haa	dhuf-	-n-	-u	haa dhuf-n-u	'Let us come'
Negative Jussive Mood						
Person	Neg.	Root	Agr.	Neg.	Inflected form	Gloss
3sgm	hin	dhuf-	-	-n	hin dhuf-i-n	'Don't let him come'
3sgf	hin	dhuf-	-	-n	hin dhuf-i-n	'Don't let her come'
3pl	hin	dhuf-	-	-n	hin dhuf-i-n	'Don't let them come'
1pl	hin	dhuf-	-	-	-	-

Table 19. Summary of Marking Jussive Mood [81]

c. Voice

Voice is a verb form that relates action of a verb with its participants (or arguments). It tells us if the subject performs or receives the action indicated by the verb. When the subject performs the action the voice is active whereas the form in which the subject receives the action is passive voice. Passive formation in Oromo is purely morphological as it is formed by adding the morpheme *-am* on transitive verbs in contrast with the unmarked active form.

Here are few examples in the table below:

Voice	Root	Marker	Inflected form	Meaning
Active	Kut-	-	kut-e	'cut'
	gurgur-	-	gurgur-e	'sold'
Passive	Kut-	-am-	kut-am-e	'was cut'
	gurgur-	-am-	gurgur-am-e	'was sold'

Table 20. Active and passive voices

4.4 Morphophonemic Processes

Morphophonemic changes are the change that takes place between the boundary of stems and inflectional or derivational suffixes [27]. In Afaan Oromo the change may be cause reduplication, assimilation, epenthesis, metathesis, deletion, of the word. Each of them is briefly discussed in the following section.

4.4.1 Reduplication

Reduplication is formed by doubling the first consonant and vowel of the verb stem and geminating the second occurrence of the initial consonant. The resulting word indicates the repetition or intensive performance of the action of the verb. Generally, if the stem starts with consonant, reduplication has the form of CV(C) + stem, where C=consonant and V= vowel. But, it has the form of V (‘) +stem if the stem starts with vowel. For example, **fiiguu=fiffiguu, utaaluu=u’utaaluu..etc**

4.4.2 Deletion

Deletion is always occurs in derivations or inflections. For example, mana ‘a house’ + -oota= manoota, nama, ‘a man’ + -icha = namicha. In verbs, deletion usually takes place in stems ending with ‘h, dh, hudhaa(‘)’. Hodh- +te=hoote Bah-+te=baate

4.4.3 Assimilation

The phonemes that come next to each other at morpheme or word boundary may take the form of the previous or next. This produces the combinations of a variety of stem-final consonants followed by t (third person singular feminine, second person singular and second person plural), n (first person plural, neutral common), s (common, causative-common singular) and so on. The change can take place between prefix and stem or stem and suffix. For example, d +s =ch for duud + sa , duucha.

4.4.4 Epenthesis

In Afaan Oromo, more than two consecutive consonants cannot occur together. When more than two consonants occur consecutively /i/ or others will be inserted between them. For instance, Elm- + -na= Elmina, and Sirb- +ta=sirbita

5 Summary

This chapter discussed morphology of different Afaan Oromo word classes. The morphophonemic processes that take place in the word boundaries have also been dealt with. The next chapter will discuss the procedures and assumptions taken in this chapter to design the morphological segmentation a valid word forms, and hence constitute the core of this study.

CHAPTER FIVE

5. METHODOLOGY

In the previous chapters we have reviewed many papers done on morphological analysis and Afaan Oromo morphology. Our proposed approach aims at learning word representations from large labeled data, generated word feature vectors are used for training. In this section we would describe architecture proposed for morphological segmentation for Afaan Oromo.

5.1 Architecture of Afaan Oromo Morphological Segmentation

Sequence to sequence (seq2seq) is general-purpose neural network architecture is used for many NLP tasks. It also called as encoder-decoder model that takes a sequence of input and generates another sequence as output. As the name suggests, encoder-decoder models consist of two parts: an encoder and a decoder. The encoder network is that part of the network that takes the input sequence and maps it to an encoded representation of the sequence. The encoded representation is then used by the decoder network to generate an output sequence. In this work, we adapt seq2seq, consisting of encoder-decoder for morphological segmentation. The architecture consists of encoder-decoder parts. The encoder part contain word2vec, encoder recurrent layer and decoder part includes word2vec, decoder recurrent layer, and decoder output layer. The following is Seq2seq based AOMS Architecture.

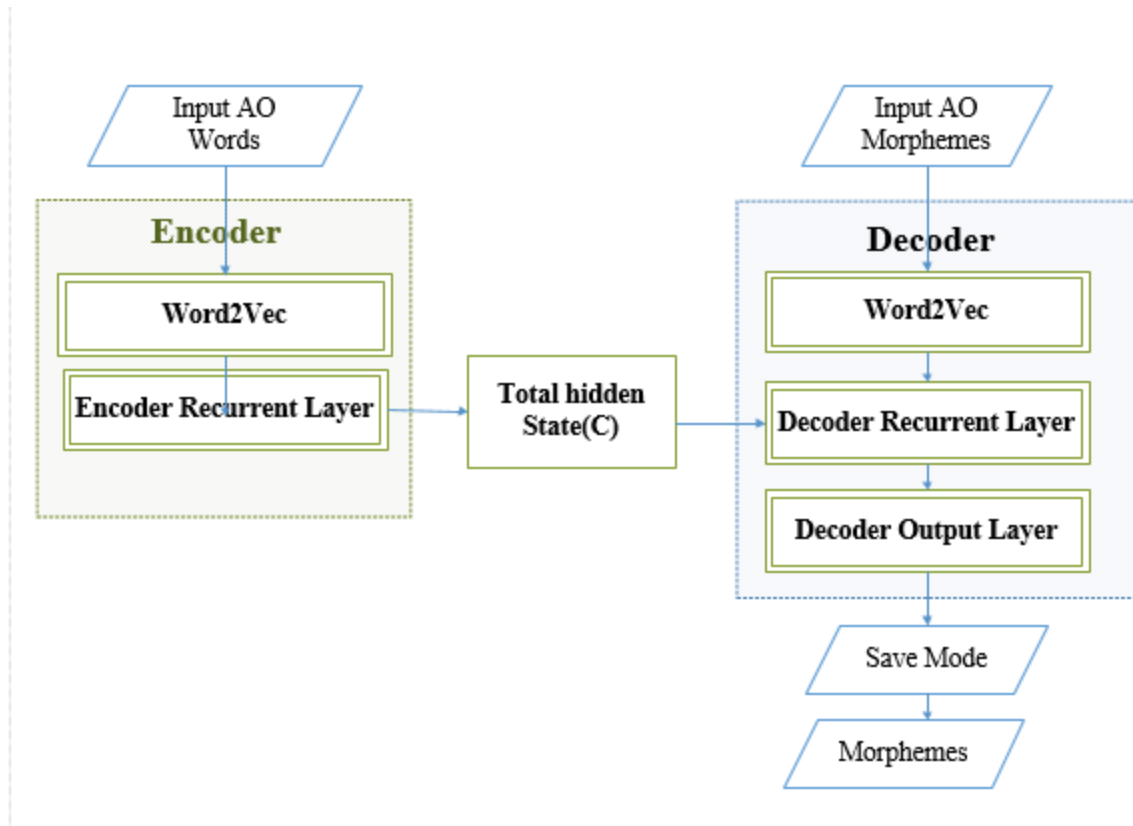


Figure 8. Architecture of AOMS

Encoder Word2vec Layer

The first layer that converts word to the vector or it reads source data, e.g. a sequence of symbols, and produces a vector containing information about this data relevant for the task. When processing i -th word in the source data the input and the output of the layer are the following [79]:

- ✓ The input is \mathbf{W}_i : i -th word
- ✓ The output is \bar{X} : the word vector which represents i -th word

$$\bar{X}_i = V(x_i) \text{-----} (5.1)$$

Encoder recurrent layer

The encoder recurrent layer generates the hidden vectors from the vectors. These vectors will be passed to a LSTM layer. LSTM layer is consists of multiple layer type of LSTM that will calculate the sequenced input. This layer will calculate the hidden state from the vector calculated in equation 5.1. For example, when processing i-th vector, the input and the output of the layer are the following:

- The input is X_i , the vector which represents the i -th word
- The output is h_i the hidden vector of the i -th position

$$h_i = f(\overline{X}_i, h_{i-1}) \text{-----} (5.2)$$

Decoder Wor2vec Layer

Decoder Wor2vec layer converts the each morphemes in the input morphemes to the vector or it reads Afaan Oromo morphemes, and produces a vector containing information about this data. When processing the j-th word in the morphemes, the output of the layer is calculated with the following equation [79]:

$$\overline{y}_j = V(y_j) \text{-----} (5.3)$$

Decoder Recurrent Layer

The decoder recurrent layer generates another hidden vectors from the vectors generated in equation 5.3 and total hidden states of input sequence in equation 5.2. In this case we must use the encoder's hidden vector of the last position as the decoder's hidden vector of first position. When processing the j-th vector, the input and the output of the layer are the following:

- The input is y_j : the vector and total hidden states of input sequence c
- The output is h_j the hidden vector of j-th position

$$h_i = h_{j-1} \text{-----} (5.4)$$

$$h_j = f(\overline{y}_j, h_{j-1}, C) \text{-----} (5.5)$$

Decoder output layer

The decoder output layer generates the probability of the morpheme from the hidden vector from which morpheme can be calculated. When processing the j -th vector, the input and the output of the layer are the following:

- The input is \mathbf{h}_j : the hidden vector of j -th position
- The output is \mathbf{p}_j the probability of generating vector \mathbf{y}_j of the j -th morphemes

$$\mathbf{p}_j = \text{softmax}(\mathbf{W} * \mathbf{h}_j + \mathbf{b}) * \mathbf{y}_j \text{-----}(5.6)$$

Generally, we used dataset consisted of word-morphemes pairs. In the case of the Encoder, each word in the input is fed into the model in a number of consecutive time-steps. At each time-step, t , the model updates a hidden vector, h , using information from the word inputted to the model at that time-step. This hidden vector works to store information about the inputted word. In this way, since no words have yet been inputted to the Encoder at time-step $t=0$, the hidden state in the Encoder starts out as an empty vector at this time-step.

At each time-step, this hidden vector takes in information from the inputted word at that time-step. This hidden vector at the final time-step inputted into the decoder. The final hidden state of the Encoder is relabeled as $t=0$. This is because this final hidden vector of the encoder becomes the initial hidden vector of the decoder. In this way, we are passing the encoded meaning of the word to the Decoder to be converted to morphemes. Before it segments a word into its morphemes, model is saved for trained data. At the end, we need the decoder to output morphemes.

Besides the Seq2seq model we have trained dataset with two other models. The first is Bidirectional LSTM model which is very similar with Seq2Seq model but, Bidirectional LSTM (BLSTM) consists of both a forward LSTM and a backward LSTM, which enables an encoding generation based on both the preceding and following tokens. The forward LSTM reads the input sequence as it is ordered (from x_1 to x_i) and calculates a sequence of forward hidden states ($-h_1, \dots, -h_m$). The backward LSTM reads the sequence in the reverse order (from x_m to x_1), resulting in a sequence of backward hidden states ($-h_1, \dots, -h_m$). Thus, when a BLSTM is used to encode an

input sequence, it generates a hidden state $-h_j$ which is a concatenation of both the forward and backward LSTMs. As in Seq2seq model there is encoder and decoder in Seq2seq BLSTM model with similar functions. The encoder part takes in an input sequence of source tokens $x = (x_1, \dots, x_m)$ and outputs a sequence of states $h = (h_1, \dots, h_m)$. The decoder is also an LSTM that computes the probability of a target sequence $y = (y_1, \dots, y_j)$ based on the hidden state h . The probability of each token in the target sequence is generated based on the recurrent state of the decoder LSTM, previous tokens in the target sequence. The training and testing process is similar to Seq2seq LSTM but to get better accuracy we have used different parameters.

To decide which RNN unit type to use, we experimented with both recurrent neural networks (RNN) and LSTM, both of which are common RNN cell variants. We found LSTM cells provided better results compared to RNN cells.

5.2 Algorithm of AOMS

- Input: *AO words with corresponding morphemes*
- Output: *Saved AOMS model*
 - Step 1: input sequences
 - Step 2: calculate word vectors
 - Step 3: define encoder and decoder as fully connected layer
 - Step 4: encoder recurrent and decoder recurrent turns their vectors to hidden vectors
 - Step 5: decoder recurrent use encoder's hidden vector and word vectors to calculate hidden state
 - Step 6: decoder output layer generates probability of output from hidden vectors
 - Step 7: Model is trained with:
 - Loss function=Categorical cross_entropy
 - Optimizer= Adam
 - Batch size=64
 - Step 8: Save model
 - Step 9: list some morphemes

5.3 Prediction Architecture

The objective of the architecture is to describe how to predict Afaan Oromo morphemes based its trained model. Preprocessing part removes unnecessary symbols and capitalization from Afaan Oromo words. Decoder segments preprocessed word based on both hidden state of input word and trained model. If the word exists in dataset, it return set of morphemes of the word, but if not it segment based on very similar word in the corpus. The following figure indicates how prediction model works.

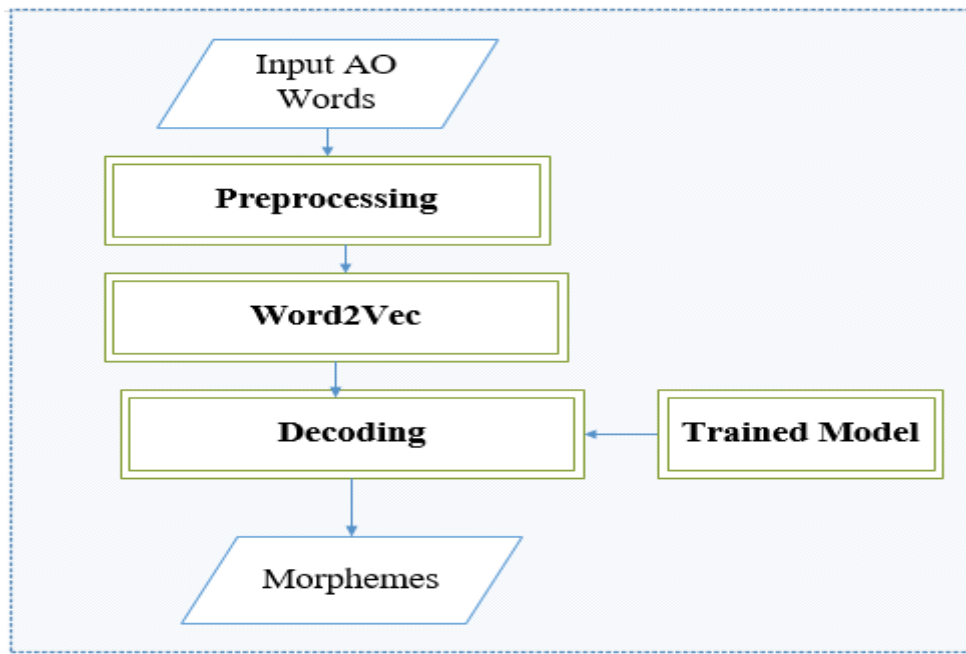


Figure 9. Prediction architecture

5.4 Algorithm to Predict Morphemes of word

- **Input:** *Afaan Oromo word*
- **Output:** *Words with its morpheme separated by* –“*sign*”
 - Step 1: input word
 - Step 2: preprocessing
 - Step 3: calculate word vector
 - Step 4: segmenting word based on trained model

CHAPTER SIX

6. EXPERIMENTS

In this chapter experimental procedure, data collection, tools, and the experimental scenarios used for evaluating our hypothesis are discussed. The model was developed and tested using a python programming language.

6.1 Data Collection

One of the main purposes of this research was constructing corpora to support the development of Afaan Oromo NLP. As we have mentioned earlier there was no publicly available morphologically segmented resource for Afaan Oromo. Therefore, we constructed the morphologically segmented corpus from books, newspapers, and online Afaan Oromo-english dictionaries. The prepared dataset consists of word and its morphemes pairs delimited by tabs. The current version of this corpus comprises over 50,200 tokens derived with its morphemes. For the purpose of the morpheme boundary detection we have used minus-sign (-) between sequence of morphemes.

6.2 Development tools

Many development tools were used in the process of this research. The tools that has been used include Tensorflow, Keras, Scikit learn and python programming.

6.2.1 Tensorflow

Tensorflow is a framework to perform computation very efficiently, Google's open source deep learning library released on November 2015. It includes C++ and Python API's. It is used as a system for building and training neural networks to detect patterns. Tensor flow considers networks as a directed graph of nodes with data flow computation and dependencies encapsulated in it [74]. Tensor flow is used as a back end for Keras library which is used for development of deep neural network models.

6.2.2 Keras

Keras is a Python machine learning library for deep learning that can run using Theano or Tensor Flow as a back end. Keras handles the way we make models, defining layers, or set up multiple

input-output models. In this level, Keras also compiles our model with loss and optimizer functions, training process with fit function. Keras doesn't handle Low-Level API such as making the computational graph, making tensors or other variables. Its main focus is enabling implementation of deep learning models as fast and easy as possible for research and development. It runs on Python 3.7 and can execute on CPUs given the underlying frameworks. It is released under the permissive MIT license [74]. Keras is used for the development of deep neural network model for proposed for AOMS system.

6.2.3 Scikit learn

Scikit-learn is a key library for the Python programming language that is typically used in machine learning projects. Scikit-learn is a Python machine learning library which includes a wide range of state-of-the-art machine learning algorithms for supervised and unsupervised problems. It is easy to use, has high performance, and contains detailed documentations [75]. Scikit learn is used to evaluate deep neural models by calculating performance metrics.

6.3 Evaluation metrics

A common approach for evaluating machine learning models is through precision (P), recall (R) and F-measure metrics as given by equation 6.1 to 6.3. The precision evaluates the percentage of number of correct boundaries found with respect to the total number of correct boundaries, and the recall measures the percentage of correct boundaries found with respect to the total number of correct boundaries. The F-measure is the harmonic mean of precision and recall and can be interpreted as their weighted average.

$$\text{Precision} = \frac{\text{number of correct boundaries found}}{\text{Total number of boundaries found}} \text{-----6.1}$$

$$\text{Recall} = \frac{\text{number of correct boundaries found}}{\text{Total number of correct boundaries found}} \text{-----6.2}$$

$$\text{Recall} = \frac{2*(P*R)}{P+R} \text{-----6.3}$$

Here the term *boundary* means the borderline between word's segments, which the algorithm succeeds or fails to discover. For example, there is one boundary in the word *si**r**b-a*. If we assume,

that algorithm segmented this word as *sir-b-a* with two boundaries, we can calculate precision as 0.5, recall as 1.0 and F-measure for this word will be 0.66. For calculating total precision, recall and F-measure simple average value are taken through them.

6.4 Experimental Setup

Laptop computer is the kind machine we have used for experiment. This machine is used for experiments done on the deep learning neural networks using windows environment. Hardware and software specifications are given in the following table.

Hardware/Software	Specification
Manufacturer	HP
Model	HP 250 G6 Notebook PC
Processor	Inter® Core™ i3-7020U CPU@2.30GHz
Memory(RAM)	4.00GB (3.92GB usable)
Operating System	Window 10

Table 21. Hardware/ Software Specification

6.5 Experimental Scenarios

To develop required model some experiments were conducted. Generally experiments can be grouped in to two. The second group of experiments was conducted to customize parameters of deep neural networks. The first group of experiments was to evaluate the performance of different models.

In order to search for the parameters that yield optimal performance, we explored hyper parameters that include batch size, dropouts, and optimizers.

Dropouts: Randomly selecting and dropping-out nodes have proven effective at mitigating overfitting and regularizing the model [76]. We applied dropout the inputs and outputs of the LSTM/BLSTM layer. The dropout probabilities are selected from a uniform distribution over the interval [0, 0.4]

Batch size: We ran tuning for the batch size of the set {64, 128, and 256}.

Optimizers and learning rate: We investigated more sophisticated algorithms such as Adam [77], and RMSProp [78], and SGD. The SGD learning rate was initialized to 0.1, a momentum of 0.9

with rate updates for every 100 epochs at a drop rate of 0.5. However, the SGD setting did not result in significant gains compared to the automatic gradient update methods.

We have used three types of deep neural networks which are Seq2Seq LSTM model, BLSTM and RNN models. By changing the recurrent layer neuron cells (LSTM & RNN) our neural net is trained with 100 iterations for each case. For evaluation train test split using 80% of corpus as training data and 20% used as test data. Categorical cross entropy objective function with different kinds of optimizers is used for training process with three models.

6.6 Training

We trained encoder-decoder model which are important for sequential data. The models used different parameters such as optimization algorithms, batch size and dropout. Optimization algorithms are used for neural network model to produce slightly better and faster results by updating model parameters such as weights and bias values. There are several commonly used optimizers; SGD, RMSprop, and Adam. SGD (Stochastic Gradient Descent) performs parameter update for each training example. It performs one update at a time. But, it requires manual tuning of learning rate which is difficult. Another optimizer is RMSprop (Root Mean Square Propagation). It is adaptive learning rate method proposed by University of Toronto professor. Adam stands for Adaptive moment estimation. It works well in practice and compares favorably to other adaptive learning-method algorithms as it converges very fast and the learning speed of the Model is quite fast and efficient and also it rectifies every problem that is faced in other optimization algorithm such as vanishing learning rates which leads to fluctuating of loss function. Therefore Adam (Adaptive moment estimation) was selected as optimizer for models we trained. Batch size refers to the number of training examples utilized in one iteration. Small batch size go through the system quickly and with less variability, which fosters faster learning. But, if batch size is big, learning rate is slow. For our models with batch size 64 we were achieved better results. To calculate error rate of the model we have used loss function. There are two most known loss function exist; Categorical cross_entropy and Binary cross_entropy. Categorical crossentropy is a loss function that is used for categorization. With categorical cross entropy, you're not limited to how many classes your model can classify. Therefore, we have used Categorical crossentropy when we have multiple classes. With binary cross entropy, you can only classify two classes. With

categorical cross entropy, you're not limited to how many classes your model can classify. It is also possible to use `categorical_crossentropy` for two classes as well.

Model Accuracy

Throughout the evaluation, the statistics used to measure the performance of the system is accuracy. Accuracy refers to the closeness of agreement between a test result and the accepted reference value. The accuracy of the system is calculated as the number of correctly generated morphemes divided by the total number of morphemes generated by the system multiplied by 100%. That is,

$$Accuracy = \frac{\text{Total number of correctly segmented morphemes}}{\text{Total Number of segmented morphemes}} * 100 \dots \dots \dots 6.4$$

the following figure indicates accuracy of the model is increasing as size of corpus increases.

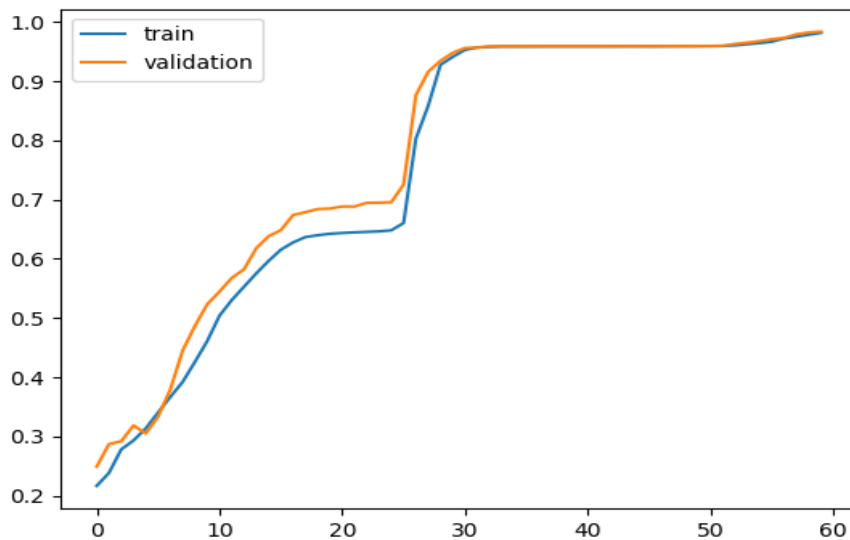


Figure 10. Model Accuracy

Model Loss

We have employed Categorical cross-entropy to calculate model loss that always used for single label classification. This is when only one category is applicable for each data point.

$$\mathbf{L}(\mathbf{y}, \bar{\mathbf{y}}) = \sum_{j=0}^M \sum_{i=0}^N (\mathbf{y}_{ij} * \log(\bar{\mathbf{y}}_{ij})) \text{-----6.5}$$

Categorical crossentropy will compare the distribution of the predictions (the activations in the output layer one for each class) with the true distribution, where the probability of the true class is set to 1 and 0 for the other classes.

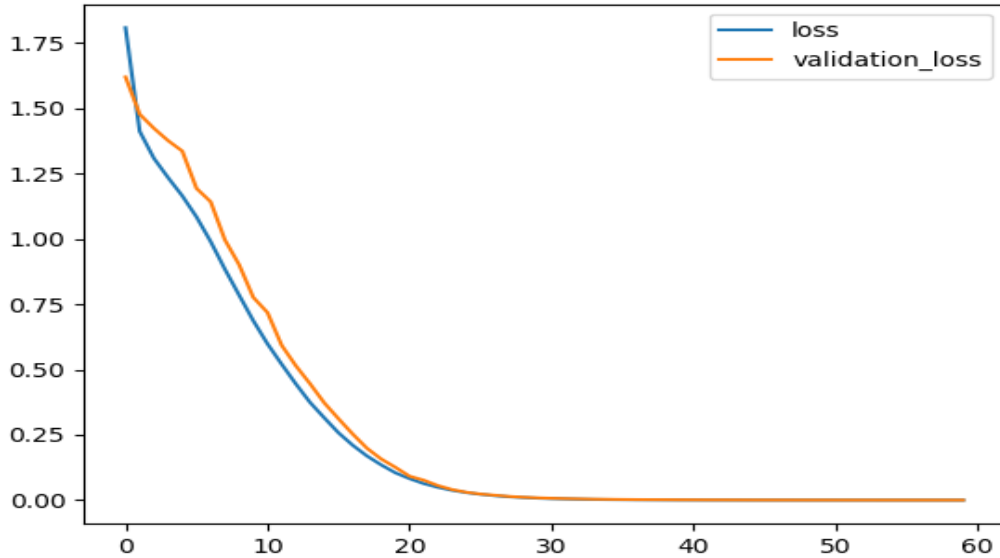


Figure 11 Model loss

The following table summarizes Loss and Accuracy of models

Model	Loss	Accuracy%
Seq2seq LSTM	1.00 X 10 ⁻¹⁰	98.8
RNN	1.00 X 10 ⁻⁸	97.5
BLSTM	1.00 X 10 ⁻⁷	96.9

Table 22. Loss and accuracy of model

6.7 Results

We trained and compared the performance of three models trained with the same morphological dataset collected. The results are summarized in the following tables.

Model	Precision	Recall	F-Score
BLSTM	97.78	97.17	97.48
Seq2seq LSTM	98.15	98.53	98.33
RNN	97.65	98.40	98

Table 23. Performance results for deep neural networks

From the above table three models have shown high performance with average F-score 97.48%, 98.33%, and 98% for BLSTM, LSTM, and RNN respectively. Seq2seq LSTM outperforms BLSTM and RNN by 0.85% and 0.33% respectively. Investigation of the results showed that all the three models were able to identify morpheme boundaries well, which is important in breaking morphemes from the original word. All of models categorical cross entropy objective function with adam optimizer is used for training process. Generally, all deep learning models were achieved more than 97% training accuracy.

There are three kinds of morphemes when we test models; *correctly segmented morphemes* are words which our model is correctly segment it to their morphemes, *incorrectly segmented morphemes* are words which our model is not correctly segment it to their morphemes and *un-segmented morphemes* are words which cannot further segmented to their morphemes. The following table summarizes them.

Model	<i>Correctly segmented morphemes</i>	<i>incorrectly segmented morphemes</i>	<i>un-segmented morphemes</i>
Seq2seq LSTM	98%	1.5%	0.5%

Table 24. Correctly Segmented, incorrectly segmented and un-segmented morphemes

The following figure is screen shot of the morphologically segmented Afaan Oromo words

```
input word : argameera
Morphemes of the word      arg-am-e-era

input word : deemtaniirtu
Morphemes of the word      deem-t-an-i-irtu

input word : faalama
Morphemes of the word      faal-am-a

input word : jalqabame
Morphemes of the word      jalqab-am-e

input word : murteerti
Morphemes of the word      mur-t-e-erti

input word : taasise
Morphemes of the word      taasis-e

input word : yaasan
Morphemes of the word      yaas-an

input word : caalchisuu
Morphemes of the word      caal-chis-uu

input word : raafama
Morphemes of the word      raaf-am-a

input word : |
```

Figure 12. Screen shot of the morphologically segmented Afaan Oromo words

Figure above shows working interface of Afaan Oromo morphological segmentation. The figure shows that the system segment word to their morpheme components correctly, but for some words such as compound words; lubbu-qabeeyyii (living things), ol-deemuu (“go up”), reduplication words such as kukkute (cut into pieces), cannot be not segmented correctly. The problem can be alleviated by training model with larger size of data.

6.8 Comparison with Baseline Experiments

As described in chapter three, to the best knowledge of researcher, morphological segmentation for Afaan Oromo has been investigated by Getacho[12] and Hornmorph [15]. The study of [12] wascauses error due to missing of suffix sequence of word. The author was segment Afaan Oromo words to their morphemes using machine learning approach which based on suffix sequences. The

experimental result the author found 94.3 F-score. In case of Hornmorph [15], the study were used a set finite state transducers were applied for morphological analyzer and generator. The author investigated morphology of three languages; Afaan Oromo, Amharic and Tigrinya. In case of Afaan Oromo, the author mainly concerned on morphological analysis, which analyze a word to its stem and its grammatical features without segmenting them into their components. In addition to this, it considered nouns and verbs. Therefore, we revised our testing part of dataset for comparison with [15]. We compared performance of our model with Hornmorph based on correctly indicated roots. As the result, out of testing dataset 65% and 98% of roots were correctly indicated by [15] and our model respectively. Finally, authors were not put the result of performance of the system for Afaan Oromo.

The current study tried to model and test using neural networks perspective of Afaan Oromo morphological segmentation using more number of datasets. The experimental result shows that the generalization performance of Seq2seq LSTM, BLSTM and RNN models are 98.33, 97.48 and 98 F-measure.

Comparison of Afaan Oromo morphological segmentation which developed using different approaches shown in the table below.

Morphological segmentation for Afaan Oromo			
Method	Size of corpus		Performance (%)
	Train	Test	
Machine learning approach (suffix sequences)	3367	500	94.3
FSTs	-	-	-
Our models	40,160	10,040	>97

Table 25. Comparison with baseline experiments

As shown in above table, our models (neural network) achieved high performance. This is due to neural networks needs requires large size of dataset to perform well. We observed that if we use small number of dataset, performance of the models would not be comparable.

6.9 Discussion

The result obtained in the experiment of deep learning neural networks is promising approach for Afaan Oromo morphological Segmentation. The models are trained on approximately 80% of the corpus and then tested on the remaining 20%. The performance of the system in terms of accuracy was determined by training the model and tuning parameter setting. The evaluation criteria used for morphological segmentation were: recall, precision, and f- score. The accuracy of the model with optimized settings are 98.33, 97.48 and 98 for Seq2seq LSTM, BLSTM and RNN.

The precision, recall and F-measure were also calculated by taking average of validation data. The results using Seq2seq LSTM with the same optimized parameters are 98.15, 98.53 and 98.33 respectively. The results obtained by BLSTM were 97.78, 97.17 and 97.48 respectively. Similarly for RNN the results obtained were 97.65, 98.40 and 98 respectively.

Number of hyper parameters used were; optimization algorithm; Adam due to it is fast than other optimizers and rectify every problem faced, Batch size=64 and Drop out = 0.1 results better performance for all models. Based these parameters, Seq2seq LSTM showed better results than others due to internal network architecture applicable for Afaan Oromo word structure. In general, by all models result achieved were greater than 97. Therefore, they are suitable for Afaan Oromo morphological segmentation.

CHAPTER SEVEN

7. CONCLUSION AND FUTURE WORKS

This chapter is conclusion of observations from our research. It also contains future works to show further researches that can be done in the future.

7.1 Conclusion

Morphological segmentation system is critical components of many NLP applications. Segmenting word into its morphemes would improve the performance of other NLP tasks that rely on it. There are different approaches have been used for segmenting word into morphemes which are rule based (linguistic approach), machine learning (statistical approach) and hybrid approach.

Several researches were done on morphological segmentation on local languages based on their rule. These researches have been used machine learning and deep learning approaches. The primary task in developing a machine learning based morphological segmentation is to learn the rules automatically which is the toughest task since the complexity of feature representation is high. Therefore, in this paper, we applied a deep learning approach for learning the rules of morpheme separation automatically.

In this research word vectors are generated that can capture relations of words and used it as a feature for our experiments. We have developed an architecture that uses word vectors as a feature by avoiding usage of manually designed features.

We have build Afaan Oromo Morphological segmentation models using deep learning neural networks. In comparison to the baseline experiments our system achieved better result. BLSTM, seq2seq LSTM, and RNN deep neural networks are also used for experiments and achieved 97.48%, 98.33%, 98% respectively. From our observations of experimental results we have concluded the following points.

- The developed models can identify the morpheme boundaries.
- Different models can use word vectors as a feature and can build AOMS model with high performance.

- From the types of BLSTM, seq2seq LSTM and RNN cells, seq2seq LSTM outperforms than other kinds models.
- We have determined suitable network parameters based accuracy of the model achieved.
- Among all models used in our experiments seq2seq LSTM model achieved the highest F-score.

Finally, from our experimental analysis we observe that lack of standard training data-set, and complexity of morphology are the most potential factors to commit errors.

7.2 Contribution of the work

The main contributions of this thesis work were summarized as follows:

- ✓ The word and its morphemes suitable for Afaan Oromo morphological segmentation system have been developed.
- ✓ The study has developed models for Afaan Oromo morphological segmentation using neural networks

7.3 Future works

This research demonstrated that automatically generated neural word vectors can be used as a feature for morphological segmentation task. It also shows that high performance morphological segmentation system can be built using word vectors. Based on our observations we recommend the following future research areas.

- We believe the deep neural network we used can be further refined by changing the network type and architecture.
- When conducting our experiments we have used small amount of corpus to build the models. By using a very large data to train models, we believe better results could be achieved. In the future, we aim to test our approach with large data.
- Our approach can be used for Afaan Oromo language, further researches can consider for other languages which have similar morphological features.

APPENDICES

Appendix A: Sample Afaan Oromo words with their morphemes

Words	Morphemes	Words	Morphemes
kutta	kut-t-a	caala	caal-a
lafa	laf-a	caalta	caal-t-a
Naannoo	naann-oo	deema	deem-a
nama	nam-a	deemta	deem-t-a
Namoota	nam-oot-a	deemteerti	deem-t-e-erti
Namootni	nam-ootni	eebba	eebb-a
Mana	man-a	eebbisa	eebb-is-a
amala	amal-a	eebbista	eebb-is-t-a
amana	aman-a	fila	fil-a
amanta	aman-t-a	fileera	fil-e-era
argata	arg-at-a	filteerta	fil-t-e-erta
bara	bar-a	filaniiru	fil-ani-iru
barnoota	bar-noot-a	gala	gal-a
bar-at-aa	bar-at-aa	galta	gal-t-a
hambaa	hamb-aa	hambisa	hamb-is-a
jaalata	jaal-at-a	kuta	kut-a
Qote	qot-e	qotan	qot-an
Fuudhe	fuudh-e	fuute	fuut-e
Barachuu	bar-ach-uu	barataa	bar-at-aa
Barattuu	bar-at-t-uu	deeme	deem-e
Deemte	deem-t-e	deemna	deem-n-a
Lafa	laf-a	lafti	laf-t-i
Oole	ool-e	oolte	ool-t-e
Raase	raas-e	raaste	raas-t-e
Uuma	uum-a	uumta	uum-t-a
Tuma	tum-a	tumta	tum-t-a
Wade	waad-e	waadde	waad-d-e
Yaada	yaad-a	yaade	yaad-e

Table 26. Sample of Dataset

Appendix B: Afaan Oromo vowels and consonants with their sounds

	Front	Central	Back
Close	i, ī		u, ū
Mid	e, ē		o, ō
Open		a, ā	

Table 27. Afaan Oromo vowels and their sounds (adopted from [67])

		Bilabial	Labio-dental	Alveolar	Postal veolar	Palatal	Velar	Glottal
Stops	voiceless plain	(p)		T			K	‘ /ʔ/
	voiceless ejective	ph /p’/		x /t’/			q /k’/	
	voiced	B		D			G	
	implosive			dh /d’/				
Fricatives	Voiceless		F	S	sh /ʃ/			H
	Voiced		(v)	(z)				
Affricates	Voiceless				ch /tʃ/			
	voiceless ejective				c / tʃ’			
	Voiced				j/ dʒ/			
Nasals		M		N	Ny /ɲ/			
Laterals				L				
Flap/trill				R				
Approximants		W			y /j/			

Table 28. Afaan Oromo consonants and their sounds (adopted from [8])

Appendix C: Sample Afaan Oromo suffixes

aa	aaf	an	aniif	aniin
arraa	atti	dhaa	dhaaf	dhaan
een	eeyyii	f	icha	ichi
ii	iif	fis	iin	iin
illee	iis	illee	irraa	irraa
ittii	ittiin	ittuu	lee	lee
n	ni	oolee	oolee	oolii
ooma	oota	ootni	oottan	rra
s	tii	tiin	tu	uma
umaa	wwan	yyuu		

References

- [1] Allen J., Natural Language Understanding. 2nd Ed. California: Redwood, Benjamin/Cummings Publishing Company, Inc, 1996.
- [2] Saranya, S. K. "Morphological analyzer for Malayalam verbs." *Unpublished M. Tech Thesis, Amrita School of Engineering, Coimbatore*, 2008.
- [3] Ak, Koray, and Olcay Taner Yildiz. "Unsupervised morphological analysis using tries." *Computer and Information Sciences II*. Springer, London, 2011.
- [4] Edward Liddy. Natural Language Processing: In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, 2001
- [5] Igor Bolshakov and Alexander Gelbukh. Computational Linguistics Models: Resources, Applications, 1st ed. 2004
- [6] Haspelmath, Martin, and Andrea Sims. *Understanding morphology*. Routledge, 2013.
- [7] Gurney, Kevin. *An introduction to neural networks*. CRC press, 2014.
- [8] Abera, N. "Long Vowels in Afaan Oromo: A Generative Approach". Master Thesis. School of Graduate Studies, Addis Ababa University, Ethiopia 1988.
- [9] Summary and Statistical Report of the 2007 Housing Census: Population Size by Age and Sex, Addis Ababa, 2008
- [10] Simon Ager, Oromo Language, *Online edition*, 2012. [Online]. Available: www.sas.upenn.edu/African_Studies/Hornet/Afaan_Oromo_19777.html [Accessed: 25-march-2019].
- [11] Gezehagn Gutema Eggi, Afaan Oromo text retrieval system, Addis Ababa University, Ethiopia, 2012
- [12] Mamo, Getachew & Melucci, Massimo & Abate, Solomon. Suffix sequences based morphological segmentation for Afaan Oromo. 2015
- [13] O. Streiter, K. P. Scannell, and M. Stuflessner, "Implementing NLP projects for non-central languages: Instructions for funding bodies, strategies for developers", *Machine Translation*, vol. 20, 2006.
- [14] Dawson, C, Practical Research Methods. New Delhi: UBS Publishers. 2002
- [15] Gasser, M.: HornMorpho: A System for Morphological Processing of Amharic, Oromo, and Tigrinya. Conference on Human Language Technology for Development, Alexandria, Egypt 2011.
- [16] Sirts, Kairit & Goldwater, Sharon, Minimally-Supervised Morphological Segmentation using Adaptor Grammars. Transactions of the Association for Computational Linguistics, 2013.

- [17] Arefyev N. V., Gratsianova T. Y. & Popov K. P. “Morphological Segmentation with Sequence to Sequence neural network”, 2018
- [18] Tedla, Yemane & Yamamoto, Kazuhide. “Morphological Segmentation with LSTM Neural Networks for Tigrinya. International Journal on Natural Language Computing”, 2018
- [19] Wang, L., Cao, Z., Xia, Y., & Melo, G.D. “Morphological Segmentation with Window LSTM Neural Networks”, 2016
- [20] Z. Yi and J. Lv, "Application of Chinese Word Segmentation Based on Linguistic Environment Analysis in Text Information Filtering System, Macau, 2009.
- [21] Stymne, S., Holmqvist, M. and Ahrenberg, L.: “Effects of Morphological Analysis in Translation between German and English”., 2008.
- [22] Carlos Gershenson, "Artificial Neural Networks for Beginners", arxiv.org
- [23] Abebe K., “Thesis: Case System in Oromo”, Addis Ababa University, 2002
- [24] Ingo Plag, “Word-formation in English”, Cambridge University Press, 2002
- [25] Andrew Carstairs-McCarthy, “An Introduction to English Morphology”: Words and Their Structure, Edinburgh University Press, 2002
- [26] Kibur Lisanu “Thesis: design and development of automatic morphological synthesizer for Amharic perfective verb forms”, Department of Information Science, AAU, 2002.
- [27] Abeshu, Abebe. “Analysis of Rule Based Approach for Afaan Oromo Automatic Morphological Synthesizer.” 2014.
- [28] Samit B. et al, “Inflectional Morphology Synthesis for Bengali Noun, Pronoun and Verb Systems”, Bangla, Dhaka, Bangladesh, March, 2005
- [29] G. A. Kiraz, “Computational Nonlinear Morphology with Emphasis on Semitic Languages”, Cambridge University Press, UK, 2001
- [30] R. Sproat, “Morphology and Computation”. MIT Press, Cambridge, MA. 1992
- [31] Cotterell, Ryan, Tim Vieira, and Hinrich Schütze. "A joint model of orthography and morphological segmentation.". 2016.
- [32] Walter Daelemans, Jakub Zavrel, Ko van der Sloot and Antal van den Bosch, TiMBL: Tilburg Memory-Based Learner version 6.3 Reference Guide: ILK Technical Report, 2010.
- [33] Michael Gasser. “Computational Morphology and the Teaching of Indigenous Languages”, School of Informatics, Indiana University, 2011

- [34] Poon, Hoifung, Colin Cherry, and Kristina Toutanova. "Unsupervised morphological segmentation with log-linear models." *Proceedings of Human Language Technologies: Association for Computational Linguistics*, 2009.
- [35] Kazakov et al. "Unsupervised Learning for Word Segmentation Rules with Genetic Algorithms and Inductive Logic Programming", 2000. Available at <http://citeseer.nj.nec.com/kuzakov00unsupervised.html>
- [36] Kaur, D. and Verma, A. "Survey on name entity recognition used machine learning algorithm", 2014
- [37] Mansouri, A., Affendey, L. S., and Mamat, A. "Named entity recognition approaches", 2008
- [38] Ruokolainen, Teemu, et al. "Supervised morphological segmentation in a low-resource learning setting using conditional random fields." *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. 2013.
- [39] Demberg, Vera. "A language-independent unsupervised model for morphological segmentation." . 2007.
- [40] Zou, Jinming, Yi Han, and Sung-Sau So. "Overview of artificial neural networks." Humana Press, 2008.
- [41] Faruqui, Manaal, et al. "Morphological inflection generation using character sequence to sequence learning." 2015.
- [42] Premjith, B., K. P. Soman, and M. Anand Kumar. "A deep learning approach for Malayalam morphological analysis at character level." *Procedia computer science*, 2018.
- [43] Shiruru, Kuldeep, "An Introduction to Artificial Neural Network". International Journal of Advance Research and Innovative Ideas in Education, 2016.
- [44] Livingstone, David. "Artificial neural networks: methods and applications". Totowa, NJ: Humana Press, 2008.
- [45] Emil M Petriu, Professor, University of Ottawa, "Neural Networks: Basics"
- [46] Liu, Yang, et al. "Topical word embeddings." , 2015.
- [47] Luong, Thang, Richard Socher, and Christopher Manning. "Better word representations with recursive neural networks for morphology." 2013.

- [48] M. Baroni, G. Dinu, and G. Kruszewski, “Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors.”, 2014.
- [49] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space”, 2013.
- [50] Y. Bengio. “Learning deep architectures for AI”. In *Foundations and Trends in Machine Learning*, 2009.
- [51] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." 2015.
- [52] Li, Shuai, et al. "Independently recurrent neural network: Building a longer and deeper rnn." 2018.
- [53] Britz, Denny. "Recurrent neural networks tutorial." *Part 1-Introduction to RNNs*, 2015.
- [54] Olah, Christopher. "Understanding lstm networks." 2015. URL <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [55] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition., 2012
- [56] Denny Britz, Anna Goldie, Minh-Thang Luong, Quoc Le, “Massive Exploration of Neural Machine Translation Architectures”, 2017, available at: arxiv.org/pdf/1703.03906.pdf
- [57] Bahdanau D., Cho K., Bengio Y., “Neural Machine Translation by Jointly Learning to Align and Translate”, 2014, available at <http://www.cl.uni-heidelberg.de/courses/ws14/deepl/BahdanauETAL14.pdf>
- [58] Tikhonov A.N., Morpheme-spelling dictionary of the Russian language, ACT, Moscow, Russia, 2008.
- [59] Kann, Katharina, Ryan Cotterell, and Hinrich Schütze. "Neural morphological analysis: Encoding-decoding canonical segments." , 2016.
- [60] Abate, Yitayal. “Morphological Analysis of Ge’ez Verbs Using Memory Based Learning.” 2014.
- [61] Wang, Linlin, et al. "Morphological segmentation with window LSTM neural networks." , 2016.
- [62] Kula, K., Varma, V. and Pingali, P. “*Evaluation of Oromo-English Cross-Language*” Technologies Research Center. Information Retrieval IIIT, Hyderabad, India (2008).
- [63]. Grage, G. and Kumsa, T Oromo Dictionary. African Studies Center, Michigan State University . 1982.
- [64]. Tesfaye G, ‘*Afaan Oromo search engine*’ Master’s thesis, School of graduate studies, Addis Ababa University, Ethiopia, 2010

- [65]. Debela T., “*Designing a Stemmer for Afaan Oromo Text: A hybrid approach*”, Master’s thesis, School of graduate studies, Addis Ababa University, Ethiopia, 2010
- [66]. Mandefro L, “*Named Entity Recognition for Afaan Oromo*”, Master’s Thesis, Department of computer Science , Addis Ababa University, Ethiopia,2010.
- [67] <https://www.mustgo.com/worldlanguages/oromo/>last accessed onMay 20, 2019
- [68] Assefa W. “Development of morphological Analyzer for Afaan Oromo”. Master Thesis, Department of Information Science, Addis AbabaUniversity, Addis Ababa. Ethiopia, 2005.
- [69]Gumii Qormaata Afaan Oromoo. Caasluga Afaan Oromoo, Jildii – 1. Komishinii Aadaaf Turizmii Oromiyaa. Finfinnee, Ethiopia, 1995 E.C.
- [70] Diriba Megersa, “Thesis: An automatic sentence parser for Oromo language using Supervised learning technique”, Department of Information Science, AAU, 2002
- [71] https://en.wikipedia.org/wiki/Oromo_languagelast accessed onJuly 15, 2019
- [72]Moti.T ,”Design of Anaphora Resolution for AfaanOromo Personal Pronoun”, St. Mary’s University, 2017
- [73] Schuster, Mike, and Kuldip K. Paliwal. "Bidirectional recurrent neural networks." 1997.
- [74] Fox, James, Yiming Zou, and Judy Qiu. "Software frameworks for deep learning at scale.", 2016.
- [75] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." 2011.
- [76] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting, 2014.
- [77] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization”, 2014.
- [78] Y. Dauphin, H. de Vries, J. Chung, and Y. Bengio, “RMSProp and equilibrated adaptive learning rates for non-convex optimization, 2015.
- [79] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation."2014.
- [80] Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. "Morfessor 2.0: Python implementation and extensions for Morfessor Baseline." 2013.
- [81] Olani, Wakweya. *Inflectional Morphology in Oromo*. Diss. Addis Ababa University, 2014.

