ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

Design of Amharic Anaphora Resolution Model

By

Temesgen Dawit

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF THE ADDIS ABABA UNIVERSITY IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN COMPUTER SCIENCE

April, 2014

ADDIS ABABA UNIVERISTY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

Design of Amharic Anaphora Resolution Model

By

Temesgen Dawit

ADVISOR:

Yaregal Assabie (PhD)

APPROVED BY

EXAMINING BOARD:

1. Dr. Yaregal Assabie, Advisor _____

2. _____ _____

3. _____ _____

Acknowledgements

First and foremost, I would like to thank my almighty God for his continuous support in my study time. You are so good and faithful to me just as your words. You always are there on my ups and downs. You made me start and complete successfully. Thank you for everything and everything yet to come.

Next, I would like to thank my advisor Dr. Yaregal Assabie for his support, advices and guidance throughout this work. You guided me starting from title selection to the completion of the thesis. I also want to thank Wondwossen Mulugeta for his supportive ideas when we had time to discuss together. I also want to thank Michael Gasser for his answer for the email I sent to him. I also would like to thank Abeba Ibrahim and Aynaddis Temesgen for the documents and supports they gave me.

I want to thank my family for their love, prayers and support in every direction I needed. Dad, Mom, Chu, Kaku, Mimi, Enush and Uchi you are blessing to me. You are role model for me. You understood me even when I failed to spend time with you. Love you all. I thank God for giving me you.

I also want to thank Tsega and Nunush for their love, appreciation and support from the beginning to the end. You facilitated everything that you can to help me. Thank you so much for everything. God bless you.

I also would like to thank Billy and all his family members for their prayers and support in everything I needed. God bless you. I also want to thank my friends Bini, Fitse, Dani, Mame, Nage, Abrish, Saha, Banchi, Titi, Yibe, Jordan and Sehin for being my side when I needed you. God bless you.

I also would like to thank Mesfine and all my classmates for their encouragements and supports. Mesfine, you helped me a lot in tagging and chunking Amharic documents and provided many supportive ideas. I really appreciate what you have done. Thank you so much. God bless you.

Finally, I want to thank all who helped me in completing this work but not mentioned here in name. Thank you so much.

Table of Contents

List of Figures

List of Tables

List of Acronyms

NLP	Natural Language Processing
AR	Anaphora Resolution
AAR	Amharic Anaphora Resolution
AI	Artificial Intelligence
POS	Part Of Speech
ASCII	American Standard for Computer Information Interchange
SERA	System for Ethiopic Representation in ASCII
NP	Noun Phrase
FBI	Federal Bureau of Investigation
UK	United Kingdom
ECB	European Central Bank
ANN	Artificial Neural Network
WIC	Walta Information Center

Abstract

Anaphora resolution is the process of finding an entity which points backward to a word or phrase that has been introduced with more descriptive phrase in the text than the entity or expression which is pointing back. An entity referring back is called *anaphor*, whereas the word or phrase being referred is called *antecedent*. Anaphora resolution is used as a component in NLP applications like machine translation, information extraction, question answering and others to increase their effectiveness. Building complete anaphora resolution systems that incorporate all linguistic information is complex and still not achieved because of the different nature of languages and their complexities. In the case of Amharic language, it is even more complex because of its rich morphology. In addition to independent anaphors, unlike other languages like English, Amharic language has anaphors embedded inside words (hidden anaphors).

In this work, we have proposed Amharic anaphora resolution model using *knowledge poor* anaphora resolution approach. The approach uses low levels of linguistic knowledge like morphology to build anaphora resolution systems avoiding the need of complex knowledge like semantic analysis, world knowledge and others. The proposed model takes Amharic texts as input and preprocesses to tag the texts with word classes and various chunks. Anaphors, both independent and hidden, and antecedents are identified from the preprocessed dataset. The model deals with both intrasentential and intersentential type of anaphors. Finally, the resolution process uses constraint and preference rules to identify the correct antecedent referred by the anaphor.

To evaluate the performance of the model, Amharic texts are collected from Walta Information Center (WIC) and Amharic Holy Bible and used as datasets. The collected dataset was divided into training and testing datasets based on 10-fold cross validation technique. Based on the collected dataset, we achieved a success rate of 81.79% for resolution of hidden anaphors whereas an accuracy of 70.91% was obtained for resolution of independent anaphors.

Keywords: Amharic anaphora resolution, knowledge poor anaphora resolution approach, hidden anaphors.

CHAPTER ONE

INTRODUCTION

1.1. Background

The word *anaphora* came from two ancient Greek words „ana" and „phora". „Ana" means back, upstream, back in an upward direction whereas „phora" means the act of carrying. So, anaphora means the act of carrying back upstream [18, 26]. It is a technique or phenomena of pointing back to an entity that has been introduced with more descriptive phrase in the text than the entity or expression which is referring back. The entity referred in the text can be anything like object, concept, individual, process, or any other thing [1]. When the referred entity is in forward direction it is called cataphora which is the reverse of anaphora. The expression which is referring back is called *anaphor*, whereas the previous expression being referred is called *antecedent*. The process of finding the antecedent for an anaphor is Anaphora Resolution [2]. The relation exists between an anaphor and antecedent increases cohesiveness of sentences because it is used frequently in both written and oral communications to avoid over repetition of terms. The correct interpretation of anaphora is vital for Natural Language Processing.

Since determining the relation existed between anaphors and antecedents is complex task, how they are related and which entity is referring to which entity in sentences is not easy to determine, resolution of anaphoric reference is one of the most challenging tasks in the field of Natural Language Processing [4]. Anaphora resolution process includes tasks like interpretation of pronouns, definite descriptions and others whose correct interpretation contributes greatly to the effectiveness of anaphora resolution process. Most of the researches performed in the area of anaphora resolution were aimed at the resolution of pronouns or one type of anaphora because it is very complex to deal with the complete anaphora resolution concepts because sometimes there are situations where human beings even face difficulty of dealing with. Anaphora resolution has a heavily interdisciplinary character. In addition to the contributions of computational linguistics, it also depends on other disciplines like logic, philosophy, psychology, neurology, communication theory and others. Though there have been many researches performed for

decades, still it needs more researches to make the problem complete which shows how complex it is. It is also considered as „AI-complete" which means solving this problem means making computers think.

Though the resolution of anaphora is complex it needs to be addressed strongly for the effectiveness of NLP applications like information retrieval, information extraction, question answering, machine translations and many others. Especially in languages like Amharic where the researches are being active, proper handling of anaphora resolution would make other NLP applications effective.

See the following examples to get clear understanding of anaphora resolution.

Examples:

1. *Abebe* loves *Meron*. *He* wants to invite *her* for dinner.
   In this example, „*he*„ *and* „*her*„ are anaphors referring to the antecedents „*Abebe*„ and „*Meron*„ respectively
2. From the beginning of March, a new airline will be operating to *Addis Ababa*. It is expected to bring new tourists to *the furst biggest Ethiopian city*.
   In this example, „*the first biggest Ethiopian city*" is anaphor referring to the antecedent „*Addis Ababa*„
3. ዛሬ *ቶዮታ መኪና* አይቻለሁ። *መኪናውም* ነጭና ማጠብ የሚያስፈልገው ነበር ።
   In this example, „*መኪናውም* „*ys* anaphor referring to the antecedent „*ቶዮታ መኪና*„
4. Wash and core *four cooking apples*. Put *them* in a fire-proof dish.
   In this example „*them*" is anaphor referring to the antecedent „*four cooking apples*„
5. *Abebe* needed a car to get *his* new job. *He* decided that *he* wanted something sporty. *Kebede* went to the Toyota dealership with *him*. *He* bought a Corolla.

   In this example, all the anaphors refer to the antecedent „*Abebe*„ The last „*He*" (He bought a Corolla) in the above example is more complex to solve than other „*He*"s. Taking the antecedent of it as both „*Abebe*„ and „*Kebede*„ seems to make correct interpretation. This kind of issues make anaphora resolution complex.

6. አበበ በሶ በላ። ከዚያም በኋላ ወደ ትምህርት ቤት ሄደ።

   In this example, the verb „ሄደ‘ refers back to the subject *‘አበበ’*. Identifying how the verb „ሄደ„ refers to the subject ‘አበበ‘ is solved with the help of morphological analizer. The morpheme information that the morphological analizers extract help identify hidden pronouns like personal pronoun „he‟ (እሱ) hidden inside the verb *‘ሄደ’*.

7. <u>ከአበበ</u> በፊት ቴአትር የሚጽፉ ከበደ በቃሉ ቸርነት ደስታ ዳዊት የሚባሉ ደራስያን ነበሩ። <u>አበበ</u> በተነሳ ጊዜ ግን <u>ከሱ</u> በፊት ወይም <u>በሱ</u> ጊዜ የነበሩትን ሁሉ በልጦ <u>አስናቃቸው</u>።

   በ1985 ዓም <u>አበበ</u> የቴአትሩ ሥራ <u>እየተስፋፋለት</u> ሄዶ ብዙ ገንዘብ <u>ስላገኘ</u> ወደ ተወለደበት አድስ አበባ ወደሚባለው ከተማ ሄዶ ውብ የሆነ ትልቅ ቤት ገዝቶ <u>አባቱ እናቱ ሚስቱና ሁለት ሴቶች ልጆቹ</u> በዚሁ ቤት ውስጥ <u>እንዲቀመጡ አደረገ</u>።

   In the above two paragraphs the underlined words represent some of the anaphors and the antecedents. For example, in the first paragraph *‘ከሱ’,’በሱ’,’አስናቃቸው‘* refer to the antecedent ‘አበበ‘ but additionally the word *‘አስናቃቸው‘* also refers to the mentioned writers**.** There are hidden anaphors in the word *‘አስናቃቸው’.* The hidden anaphors in it are „he„ (እሱ) and „they„ (እነርሱ**)**.

   To identify these kinds of hidden anaphors it needs morphological analyzer which makes Amharic anaphora resolution process complex than other languages like English.

1.2. Statement of the problem

It could be easy for humans to identify which anaphora identifies which word or antecedent in a sentence or to extract hidden anaphors from sentences but it is hard for machines to correctly do so. When it comes to Amharic language it becomes even harder because it is a morphologically rich language like other Semitic languages.

Studies show that many entities or words in sentences are referred by anaphors, yet they are ignored in many NLP applications like question answering, machine translation, information extraction, opinion mining, text summarization and other many applications. But the use of anaphora resolution in these applications improves their performance [5].

Currently, NLP for Amharic language is being one of the major research areas in Ethiopia but there is nothing so far done in the area of anaphora resolution as to the best knowledge of the researcher.

1.3. Objectives

The general and specific objectives of this work are stated below

1.3.1. General Objective

The general objective of this research work is to design a model for Amharic anaphora resolution

1.3.2. Specific Objectives

The specific objectives are:

- Review literatures in the area of anaphora resolution and different approaches needed to solve it.
- Study the nature and characteristics of Amharic language
- Collect Amharic documents to be used as dataset
- Develop a model for Amharic anaphora resolution
- Develop a prototype using the model developed
- Evaluate performance of the model

1.4. Methods

The methods followed to achieve the objectives of this work are described in the following sections

Literature review

Literature review is one of the most important tasks or methods needed for the successful completion of research papers. So, the researcher reviewed different kind of anaphora resolution related books, articles, conference proceeding papers, thesis documents, and others many to get clear understanding on the subject area. Different kind of existing approaches to solve the anaphora are also reviewed. Moreover, the already designed anaphora resolution systems for other languages are reviewed. In addition to all this, to understand the nature and structure of Amharic language, books, articles and others which are useful for this understanding are reviewed.

Data Collection

Initially our plan was to collect Amharic documents which are POS tagged. But, we didn‟t get POS tagged Amharic texts as we planned. So, in addition to collecting tagged documents, we annotated some Amharic texts manually. Preparing and collecting dataset for this work was very challenging task. The collected and tagged documents are used for both training and testing datasets.

Tools

Python 3.2 programming language is used to develop a prototype for Anaphora Resolution Model.

Hornmorpho is used to generate morphological information of collected Amharic texts (i.e. to generate morphemes of those texts). It is a Python program that analyzes Amharic, Oromo, and Tigrinya words into their constituent morphemes (meaningful parts) and generates words, given a root or stem and a representation of the word‟s grammatical structure [6].

Amharic Chunker is used to chunk POS tagged Amharic documents. It is a system developed to chunk POS tagged Amharic documents [19]. Initially our plan was to chunk all POS tagged documents using this system; but it needs some modification to chunk and also manually preprocessing input data. As a result, we chunked some Amharic texts manually.

SERA is used to Romanize Ethiopic scripts. It is a system to convert Ethiopic script into the seven bit American Standard for Computer Information Interchange (ASCII) [36].

Prototype Development

Prototype is developed to test effectiveness the model using python 3.2 programming language.

Testing

Performance of the model is tested by giving preprocessed Amharic texts and checking whether anaphors in the document are matched to their antecedents correctly or not.

The following methods or formulas were used to test the model:

$$Sucess\ rate = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ all\ anaphors} \quad (1.1)$$

$$Critical\ sucess\ \text{rate} = \frac{Number\ of\ correctly\ resolved\ anaphors}{Number\ of\ anaphors\ with\ more\ than\ one\ antecedent} \quad (1.2)$$

In the formula (1.2) above, Number of anaphors with more than one antecedent are number of anaphors left after morphological is filter applied.

1.5. Scope and Limitation

The scope of this research is limited to the following stated points:

- Cataphora is not considered.
- The model will be applied only for text or written documents.
- The model assumes full sentences which are grammatically, syntactically, and semantically correct.
- The model will be applied only for Amharic pronominal anaphora types specifically personal pronouns.

1.6. Significance of the study

Anaphora is used to maintain cohesiveness of sentences i.e. they maintain flow of sentences. As a result, they are important to determine the meaning of sentences. When they are not treated properly they could even change the meaning or interpretation of a sentence.

Nowadays, anaphora resolution is being addressed in many NLP applications because it improves their performance when treated properly [5, 7, 8, 19, 21, 27]. The following listed NLP applications are some of the applications where anaphora resolution can be applied to improve their performance:

- Machine Translation
- Information extraction

- Text summarization
- Question answering
- Opinion mining
- Dialogue systems

Moreover, this research work will motivate researchers to work on this area for further study.

1.7. Thesis Organization

The rest of the thesis is organized as follows. Chapter 2 discusses the process of anaphora resolution, different forms of anaphora, background history on anaphora resolution, anaphora resolution approaches, introduction of Amharic language and other related concepts. Chapter 3 discusses works which are related to anaphora resolution. Since there is no work in the area of anaphora resolution for Amharic language, to the best knowledge of the researcher, related works in other languages are discussed. Chapter 4 discusses the design and implementation of anaphora resolution model. It discusses about architecture of Amharic anaphora resolution, prototype development and algorithm used. Chapter 5 discusses about dataset preparation, testing methods used and performance result of the model. Conclusions and recommendations of this work are presented in chapter 6. References are given at the end.

CHAPTER TWO

LITERATURE REVIEW

Researcch in the area of anaphora resolution is started in 1960s with the start of research in the area of artificial intelligence. Main focus of the projects designed those days was making systems intelligent enough to accept instructions from humans in natural language for simple tasks. To make that objective achievable, addressing of anaphora resolution, especially pronoun resolution was necessary. The systems called STUDENT and SHRDLU were very famous among the systems designed at that time. The system called STUDENT was designed to give answers to word problems found in high school algebra lessons whereas SHRDLU was designed to give commands to a robot which was capable of moving objects. The approach that those systems followed was limited to a number of heuristic rules while lacking deeper theoretical knowledge in the area [18, 26, 40]. This chapter discusses about the different forms or types of anaphora, anaphora resolution process, anaphora resolution approaches and other related issues. Moreover, it gives brief introduction on Amharic language.

2.1. Forms of anaphora

Anaphora can be divided into pronominal anaphora, lexical noun phrase anaphora, noun anaphora, verb anaphora, adverb anaphora or zero anaphora based on the form of the anaphor or syntactic category of the anaphor. It can also be classified into intrasentential or intersentential based on the location of the anaphor [4, 21, 27, 28]. The different types of anaphora mentioned above are discussed in the following sections.

2.1.1. Pronominal anaphora

Pronominal anaphora is type of anaphora that is used in many research papers and as a result it is the most common and known type of anaphora. It occurs when the anaphors are pronouns [2, 4, 18, 21, 27, 28].

Example,

Today I met Betel and her boyfriend.

In this example, „*her'* is the anaphor and „*Betel'* is its antecedent.
Personal, possessive and demonstrative pronouns both singular and plural are categorized under pronominal anaphora [21].

2.1.2. Lexical noun phrase anaphora

This type of anaphora occurs when the anaphor is categorized as definite noun phrase while the antecedent is proper name [2, 4, 21, 27, 28].

Example,

> Hailemariam welcomes the nomination of Daniel to the vice-presidency of the ECB. The Prime Minister of Ethiopia said "Ethiopia would be very well represented by the governor of the National Bank of Ethiopia, Daniel Berhan, were to be chosen for the post of vice-president of the European Central Bank".

In the above example the definite noun phrase „*The Prime Minister of Ethiopia'* is the anaphor whereas the proper name „*Hailemariam'* is its antecedent. This form of anaphora increases cohesiveness of the sentence by adding more information to the sentences. Lexical noun phrase anaphora may appear in several forms.

- when anaphors and antecedents have the same head [28]:

  Example,

  Today I ate enjera at breakfast. That enjera was really tasty.

In the above example, „*That enjera'* is the anaphor for the antecedent „*enjera'*.

- When antecedents and anaphors are different in words but synonyms [28].

  Example,

  The police ordered the *bus* to stop and asked the driver to leave the vehicle.

In the above example, „bus" and „vehicle" are synonyms. The anaphor „vehicle" refers to the antecedent „bus"

- When the anaphor and antecedent have generalization or hypernym relationship [28]:

  Example,

  Addis Ababa won the tourists choice awards for „fastest growing city in Africa". The Ethiopian capital has received the award this year.

In example above, the more generalized word „Addis Ababa" is referred by „The Ethiopian capital" i.e. „The Ethiopian capital" which is the anaphor referring back describes the more specific or general word „Addis Ababa".

- When the anaphor and antecedent have specialization or hyponymy relationship [28]:

  Example,

  Ethiopian woman wins gold medal in athletics. Tirunesh Dibaba beat at the final the Keniyan contestant Eunice Jepkorir.

In the above example, more general or specific word „Tirunesh Dibaba" refers back to the word „Ethiopian woman", which means the word, anaphor in this case, that needs description or explanation about itself is referring to the antecedent „Ethiopian woman" that gives more information than the anaphor.

2.1.3. Noun anaphora

This type of anaphora occurs when there is a relation between anaphor and antecedent as name and noun phrase [27, 28].

Example,

The dog is sleeping. Anbes, as the owners call him, enjoys long naps.

In the above example, a name is anaphor for the sentence. „Anbes" refers to the antecedent „The dog".

2.1.4. Verb anaphora

This type of anaphora occurs when anaphor is verb while antecedent is verb or verb phrase [4, 21, 27, 28].

Example,

As the mother told the child to stop walking it did so.

In the example above, the verb anaphor „did" refer to the verb phrase antecedent „stop walking".

2.1.5. Adverb anaphora

This type of anaphora occurs when the anaphor is adverb [21, 27, 28].

Example:

I was born in Addis Ababa and live here since ever.

In the example above, the anaphor „here" refers to the antecedent „Addis Ababa".

2.1.6. Zero anaphora

This type of anaphora occurs when anaphor referring to antecedent is zeroed which is to mean that it occurs when anaphor is not mentioned explicitly in sentences. Some of the most common types of zero anaphora are presented as follows [2, 4, 21, 27, 28].

- Zero pronominal anaphora [28]:
  This type of anaphora occurs when pronominal anaphors are omitted from sentences as the following example.

  Example,

  Bereket is sleepy. (He) Diddn"t sleep all night.

- Zero noun anaphora [28]:
  This type of anaphora occurs when the head noun of Noun phrase is omitted in the sentence.

Example:

There were many lead guitars in the shop before one hour, now there are none.

- Verb phrase zero anaphora [28]:
  This type of anaphora occurs when the anaphor, verb phrase in this case, is omitted from the sentence.

Example:

Cherinet wanted to go to Diredawa, but Misgana did not.

2.1.7. Intrasentential and intersentential anaphora

As it is discussed in Section 2.1 anaphors are also classified into intrasentential or intersentential based on the location of the antecedents. If the anaphor and its antecedent occur in the same sentence it is called intrasentential anaphora whereas if they occur in different sentences it is called intersentential anaphora [2, 18, 21, 27, 28].

Example:

1. *Abebe* loves *Meron*. *He* wants to invite *her* for dinner.
2. Today I met Betel and her boyfriend.

From the above examples, the first example shows intersentential type of anaphora because the anaphors and antecedents are in two different sentences, whereas the second example shows intrasentential type of anaphora because the anaphor and antecedent found in the same sentence.

2.2. The process of anaphora resolution

The study of anaphora is a complex phenomenon which needs the involvement of different areas of linguistics such as morphology, syntax, semantics and discourse. However, it goes beyond the area of linguistics and attracted the attention of researchers from different disciplines such as psychology and computer science. It also needs fundamental issues of world knowledge representation and reasoning. Resolving anaphora automatically is called *anaphora resolution* [27].

Government and Binding Theory, Centering Theory (CT), Rhetorical Structural Theory (RST) and Discourse Representation Theory among others are the theories that laid a strong ground for the growth of anaphoric related issues [40]. Government and Binding Theory revolutionized the way of investigating language and gave many valuable insights to the issues of anaphoric reference. The famous principles of this theory are known as Principles A, B and C [40]. Centering Theory is a complex theory that models issues related to prominence of discourse objects. Just like binding theory it makes various claims about anaphoric expressions and their referential properties. But it also studies anaphoric expressions and their referential properties from a different perspective through investigating local textual coherence. The algorithm called as *BFP* is probably the most known algorithm utilized the concept of Centering Theory to interpret pronominal anaphora [40]. Discourse Representation Theory (DRT) is another theory which helped the growth of anaphora resolution by providing rules. This theory provides rules to solve anaphora by giving context information about the sentences. Context information about sentences is captured because Discourse Representation Structures which is built and updated algorithmically by rules proposed by this theory and based on syntactic representation of sentences captures all information about entities mentioned in the sentence and the relations among them [40]. Sources of knowledge required in the process of resolving anaphors are presented as follows [27].

- **Morphological and lexical knowledge:** are among the most important parts to resolve anaphora correctly. Part-of-speech, gender, number and person information are categorized in this part. They give us information necessary to identify the pronouns, which are among the types of anaphors, easily. This part is also needed to minimize the candidate lists for the anaphors by discarding candidates which don"t fulfill the criteria to be selected as correct antecedent of the anaphors like disagreement of anaphors with antecedent in number, gender and person.

- **Syntactic knowledge**: information about constituents like NPs, clauses, sentences can be identified using syntactic knowledge of the language.

- **Semantic knowledge:** this knowledge helps to validate whether the anaphors referring to the antecedent is semantically correct or not i.e., it helps to check whether the anaphor-antecedent relation makes sense or not.

- **Discourse knowledge**: anaphora is has discourse phenomenon because it contributes to the cohesion and coherence of discourse. As a result, knowledge about discourse structure is important to resolve the anaphora.

- **Real-world (commonsense) knowledge:** This is the most difficult of the knowledge"s required to apply and resolve anaphors. For example, it needs solving "the FBI", "the Pope" and "the UK" automatically when they are found in sentences which needs real world knowledge to solve it.

Depending on the inherent characteristics of languages, AR may pass through several steps. However, the steps listed below are three main steps needed to resolve anaphors [27].

i. **Identification of anaphors**: all anaphors found in dataset are identified by this step (i.e. the resolution of anaphors starts with identifying which anaphors to solve). Tokenizer, part of speech tagger and chunker are used to perform this task.

ii. **Identifying the location of candidates:** This stage is used to identify potential antecedents from the dataset. It may need going backward or forward from the position of anaphors. Most of the time search scope is limited to like 2 or 3 sentences because of the performance issues.

iii. **Selection of an antecedent from the set of candidates**: This stage is used to propose the correct antecedent from a set of candidate antecedent if it is found, or propose nothing if the correct antecedent is not found.

Tokenization, part-of-speech tagging and determination of noun phrase structure are the minimum level of tasks needed to process the input text to AR systems [27]. These tasks make the identification of anaphors and antecedents easy. Mostly noun phrases are the most commonly focused areas used as antecedents to solve the anaphors because others like verb phrases,

paragraphs, sentences as antecedents to solve anaphors is very complicated task. Within the specified scope limit initially all noun phrases can be potential candidate antecedents [18]. So, the only noun phrase or noun which best matches the anaphor in question is selected from the candidate antecedents.

2.3. Constraint and preference rules

Antecedent indicators or factors is a general name given for constraint and preference rules. They are crucial for the resolution of anaphora [18, 12, 27]. In the below sections we will discuss some of those rules. Constraint and preference rules mentioned below are not exhaustive; we mentioned only some of the commonly used rules. Some of the rules are general and can be applied to all languages whereas some are not applicable to all language. Both rules are discussed in the following sections.

Constraint rules: are called *eliminative* because they must need to be satisfied for any antecedent to be considered as candidate antecedent for the anaphora resolution. The antecedents that can‟t satisfy constraint rules are discarded and will not pass to the next step of anaphora resolution. So, when the antecedents couldn‟t satisfy the constraint rules they are taken as they can‟t be correct antecedent of the anaphors and discarded. Their strength is measured on how well they filter the antecedents. Gender, number and person agreement, c-command constraints, selectional restrictions are common constraint rules used in most researches [18, 21].

Number, gender, and person agreement: are constraint rules used in most research papers. Noun phrases and anaphors usually match in number (singular or plural), gender (male or female) and person (first person, second person, third person). If anaphor is singular then the number constraint expects antecedent to be singular as well. Gender and person agreement constraints are also treated in the same manner [7, 18, 21, 26, 28]. But, there are exceptions to these cases. Sometimes semantically correct sentences can have plural anaphors referring to singular antecedents; same for gender and number. Exceptional cases for Amharic language are presented in chapter 5 of this document.

Selectional Restrictions: this constraint rule is categorized under semantic knowledge‟s required to resolve anaphora. It expects anaphor and antecedent be matched semantically. See the following examples for more clarification [21, 28].

Example:

1. Abebe ate a motor-bike
2. I will eat my hat.

Both of the examples provided above are semantically wrong, because a human being can‟t eat either motor-bike or hat. So the selectional restriction checks such kind of cases.

Preference rules: are called *preferential* because they don‟t discard any antecedents. They give more preference to antecedents when they are satisfied. Unlike the constraint rules they are not mandatory. They are applied on antecedents passed constraint rules and when a certain anaphor has more than one antecedent as a choice. Antecedents satisfied preference rules get some value whereas those didn‟t satisfy the rules given nothing [18, 21]. Some of the preference rules are definiteness, giveness, recency, frequency of mention, indicating verbs.

Definiteness: in English language, noun phrases or antecedents are said to be definite when the head noun is modified by a definite article, or by demonstrative or possessive pronouns. Definite noun phrases in previous sentences are more likely antecedents of pronominal anaphors than indefinite ones [7, 12].

Givenness: in English language, nouns phrases or antecedents are said to be „given‟ when they are not new to the current discourse or paragraph (i.e. when they are already known). Noun phrases in previous sentences representing the „given information‟ are assumed to be good candidates [2, 12].

Indicating verbs: this preference rule may not apply to all languages. In English language, the first Noun phrase identified after the set of verbs presented below is assumed to be good candidates for anaphors [2, 12].

Verb_set = {discuss, present, illustrate, identify, summarize, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyze, synthesize, study, survey, deal, cover}

Recency: this preference rule checks whether the antecedent is closer to the anaphor in question or not. It can be measured by counting number of words or sentences between antecedent and anaphor. A candidate that is closer to the anaphor receives higher preference [7, 21, 31].

Frequency of mention: this preference rule checks how many time a candidate antecedent is mentioned in a given discourse. If a word or antecedent is mentioned frequently in a given discourse then it has more probability to be selected as correct antecedent because it is given higher preference [2, 7, 12, 21].

2.4. Anaphora Resolution Approaches

Researches performed in the area of anaphora resolution in the beginning gave good theoretical results and as a result it was believed that the research in this area would take few years to implement systems having every aspect of anaphora resolutions. But, after years of intensive research it is accepted that building automatic anaphora resolution systems having all anaphora related aspects is tough to accomplish. The reason is because the formalization of world knowledge, various semantic issues and other issues which are important for the implementation of successful anaphora resolution systems are still far from being achieved. As a result, in 1990s most researches were shifted to exploiting lower-level information like morphology and syntax from the higher level information which are complex. So, based on the level of information required to resolve anaphora, two broad categories have emerged: knowledge-rich and knowledge-poor approaches. Both approaches are discussed in the following sub sections [12, 18, 21, 29, 40].

2.4.1. Knowledge-rich Approaches

Knowledge rich anaphora resolution approaches are approaches that employ linguistic and domain knowledge in great detail for anaphora resolution process starting from the

morphological knowledge to the very high level knowledge like world knowledge. These approaches are rule based. Knowledge‟s like morphological, syntactic, semantic, discourse, domain knowledge are needed to deal with anaphora resolution using knowledge rich approaches. Most traditional approaches or systems are knowledge rich [29].

Knowledge-based anaphora resolution approaches can be divided in to four: discourse oriented approaches, factor based approaches, syntax based approaches and heuristic based approaches [15].

Discourse oriented approaches assume that some entities in the discourse are more central or focus than others which shows relations existing between texts in discourse. When there are some texts which are more center or in focus than others in discourse that gives some relations between the texts found in discourse then it is easy to perform the anaphora resolution process because texts which are more center have more probability than others to be selected as the correct antecedent. Centering theory is a theory used to track the center of discourse. It is used to interpret pronouns by modeling the local coherence of a discourse. Factor based approaches are another category of knowledge-based approaches that uses constraints to remove wrong antecedents, and preferences to rank candidates which are satisfied by constraint rules. Since it is knowledge-rich approach, the constraint and preference rules used include morphological, lexical, syntactic, semantic and pragmatic information. In general, it depends on some set of factors to resolve anaphora. Syntax based approaches are still another part of knowledge based approaches. It is an approach which totally depends on syntactic and morphological information to perform anaphora resolution process. It uses parse trees to search potential antecedents. This approach is limited to pronoun resolution only because syntactic structures may not handle other types of NPs. Heuristic based approaches are another categorization of knowledge based approaches which use a set of heuristics or rules to select the correct antecedent of the anaphora to be solved. These approaches require the researcher to produce a set of heuristics to resolve the anaphora. To sum up, since knowledge based anaphora resolution needs complex syntactic, semantic, discourse and world knowledge analysis, using it as an approach to solve anaphora is labor intensive and time consuming.

2.4.2. Knowledge-poor Approaches

Knowledge poor approaches are approaches formulated to avoid the complex syntactic, semantic, discourse and world knowledge‟s used in anaphora resolution process. Since knowledge based approaches are labor intensive and time consuming plus computationally expensive, knowledge poor anaphora resolutions are formulated to solve these complexities. It is a result of high need for inexpensive solutions to satisfy the need of NLP systems in practical way. The development of NLP tools like POS tagger and other tools facilitated for the invention of knowledge poor approach. This approach makes use of part of speech tagger, noun phrase rules and then applies antecedent indicators which are constraint and preference rules on a set of potential antecedents. Knowledge poor anaphora resolution approach depends on a set of constraint and preference rules [12, 18].

Anaphora resolution approaches that showed the practicality of knowledge poor anaphora resolution approach include: Kennedy and Boguraev‟s approach without a parser [18], robust, knowledge poor approach [12, 18], CoGNIAC [16, 18], collocation patterns-based approach [29, 30], machine learning approaches [13, 29, 30], probabilistic approach [29, 30]. In the following paragraphs we will discuss of some of the works briefly.

Kennedy and Bogurav‟s approach without a parser is an anaphora resolution approach which is categorized under knowledge-poor anaphora resolution approach. It works by taking the output of POS tagger as input. This approach uses salience weights which are called antecedent indicators to identify the correct antecedent. As other knowledge-poor anaphora resolutions do, it takes the candidate with highest salience weight as the correct antecedent for the anaphor. When there is complexity of choosing correct antecedent like if two candidate antecedents have same salience weight then the candidate closest to the anaphor is selected as correct antecedent.

Robust, knowledge-poor approach is categorized as knowledge-poor anaphora resolution approach as its name implies. This approach works by taking POS tagger output as input. It identifies noun phrases preceding the anaphor within 2 sentences distance as antecedent, applies constraint rules (gender and number agreement) on the identified antecedents to check whether

they agree with the anaphor or not. The antecedents passed the constraint rules are then assigned preference rule values. The antecedent having highest preference value is selected as the actual antecedent being referred by the anaphor. Best collocation pattern score, candidate with higher score for indicating verb, and finally most recent candidates are the order of priority level when there are more than two antecedents left after the application of all preference rules.

CogNIAC: is a system developed to resolve pronouns with limited knowledge and linguistic resources. It is called *A High Precision Pronoun Resolution Engine*. The main assumption of it is that there is a subclass of anaphora that does not require general purpose reasoning. Inputs to the system are pre-processed by NLP tools like POS tagger. It is built on the rules listed below.

i. Unique in discourse: this rule applies when there is only one possible antecedent in the anaphora resolution scope defined. If there is only one possible antecedent in the scope then that antecedent is taken as correct antecedent.
ii. Reflexive: this rule applies when nearest possible antecedent in the anaphora resolution scope defined is reflexive pronoun. In that case the correct antecedent is the antecedent which is more recent to the reflexive pronoun.
iii. Unique in current and prior: this rule applies when there is only one antecedent found in the previous and the current sentence. In that case that antecedent is taken as correct antecedent.
iv. Possessive pronoun: this rule applies when the anaphora is possessive pronoun and when there is only one antecedent in the previous sentence. In that case that antecedent is selected as correct antecedent
v. Unique current sentence: this rule applies when there is only one antecedent in the current sentence. In that case that antecedent is selected as correct antecedent
vi. The sixth rule applies when the subject of previous sentence contains only one antecedent and when anaphor is the subject of the current sentence. If this is true then that antecedent is taken as correct antecedent.

CoGNIAC resolves pronouns from left to right in the text. For the pronouns identified, the rules mentioned above are applied according the order presented. If the antecedent is found by applying the first rule then process stops i.e. the rules next to it are not be applied. The next rules

are applied in their order of presentation whenever the prior rules are not true. If all the presented rules couldn‟t found correct antecedent then it is left unresolved.

2.5. The Amharic Language

Amharic language, which is categorized under Semitic languages family, is a national language of Ethiopia (i.e. it is official working language of the federal democratic republic of Ethiopia). It is the second most spoken language among Semitic language families in the world, next to Arabic. Since it is official working language of Ethiopia, there are number documents produced in Amharic like formal letters used for communication between different organizations or within a given organization, working manuals and others. Though the language is widely spoken or used as working language on a country having more than 90 million people, it is still categorized as „under resourced‟ language like many other African languages [32, 33, 37, 39].

Amharic language has its own writing system having scripts originated from the Ge‟ez alphabet. The writing system is rich syllable patters or fidels having atleast 41 consonant classes each having 7, 8, 12 or 13 forms. It is written from left to right. Amharic morphology and personal pronouns are discussed in the following sub sections [36].

2.5.1. Amharic Morphology

Amharic language is morphologically compex due to the semetic languages nature. Unlike languages like English, gender, number, definiteness, prepositions and others information are attached to Amharic Nouns and adjectives that resulted in complex morphology of the language [23, 41]. For example, from the noun መኪና (mekina/car), the following words are generated through inflection and affixation: መኪናዎች (mekinawoc/cars), መኪናው (mekinaw/ the car {masculine}/his car), መኪናየ (mekinaye/my car), መኪናየን (mekinayen/my car {objective case}), መኪናሽ (mekinax/your {feminine} car), ለመኪና (lemekina/for car), ከመኪና (kemekina/ from car), etc. It is also possible to generate the following words from the adjective ትንሽ (tnx/small): ትንሹ (tnxu / small, {definite} {masculine} {singular}), ትንሾች (tnxoc/ small {plural}), ትንሾቹ (tnxocu/ small {definite} {plural}), etc.

Inflections and derivations of Amharic verb are even more complex than that of Amharic nouns and adjectives. It is because many verbs in surface forms are generated from a single verbal stem, and many stems in turn are generated from a single verbal root.

Combination of person, gender, number, case, tense/aspect, and other information are extracted from Amharic verbs resulting in thousands of words from a single verbal root [1], [19]. As a result, a single word may represent a complete sentence cosutructed with subject, verb and object. For example, አልፈልግም (ysebreNal/ I don‟t want) is a complete sentence.

2.5.2. Amharic Personal Pronouns

Pronouns in Amharic language can be classified into independent pronouns and embedded pronouns based on how they exist in sentences [23]. Independent pronouns are pronouns which can be found independently in sentences like personal pronoun እርሱ in the example below.

Example,

መልእክተኞቹ ወደ ንጉሱ ተመለሱ ። እርሱ ግን ለምን እንደተመለሱ ጠየቃቸው ።

whereas embedded pronouns are pronouns bounded as affixes to words as shown in the following example.

Example,

ካሳ የገዛው የትናንቱ በግ ታረደ

In the example above the verb ታረደ shows personal pronoun እርሱ embedded in it. For the detail description of embedded pronouns see Chapter 1 and Chapter 4. Table 2.1 shows list of independent Amharic personal pronouns.

Table 2.1 List of Amharic independent personal pronouns (Adapted from [23])

Independent personal pronouns in Amharic	Personal Pronouns in English	Independent personal pronouns Romanized	Person Information	Number information	Gender information
እኔ	I	InE	1st person	Singular	Male/Female
አንተ	You	ante	2nd person		Male
አንቺ	You	anci	2nd person		Female
እሱ	He	Isu	3rd person		Male
እሷ	She	IsWa	3rd person		Female
እኛ	We	INa	1st person	Plural	Male/Female
እናንተ	You	Inante	2nd person		Male/Female
እነሱ	They	Inesu	3rd person		Male/Female
እርስዎ	You	Irswo	2nd person	Singular	Male/Female
እሳቸው	He/She	Isacew	3rd person		Male/Female

Total number of independent personal pronouns in Amharic is 8 but there are two additional personal pronouns (እርስዎ and እሳቸው) to show respect or politeness as shown in the table above.

CHAPTER THREE

RELATED WORK

As discussed on chapter 2, research in the area of anaphora resolution is started in 1960ʺs. Starting from that time many researchers discovered methods and other related concepts which are being used as a reference even today. Many research papers are published starting from then in languages other than Amharic. To the best knowledge of the researcher so far no research is done in the area of anaphora resolution for Amharic language. This chapter presents some of the works related to anaphora resolution done in languages other than Amharic.

3.1. A knowledge-poor pronoun resolution system for Turkish

Dilek Kucuk [21] used knowledge poor anaphora resolution approach to resolve Turkish pronouns. The work focused on third person personal pronouns and possessive pronouns referring to proper person names as antecedent. Since it is knowledge-poor approach the researcher used constraint and preference rules. Constraint rules used in this work are number agreement, reflexive pronoun constraint, personal pronoun constraint, and selectional restrictions. Preference rules used are quoted/unquoted text preference, nominal predicate preference, repetition preference, punctuation preference, antecedent of zero pronoun preference, and syntactic parallelism preference.

The researcher used empirical analysis to determine the search scope of antecedents, and constraint and preference rules. Some of the constraint and preference rules used in this work are the contributions of the researcher which were not used before anywhere. The value of preference rules are determined by Artificial Neural Network (ANN) as presented on Table 3.1. The search scope of antecedents defined in this work after the empirical analysis is four sentences including the current sentence.

Table 3.1 list of preference rules and their values used for Turkish language [21]

Preference rules	**Scores**
Quoted/Unquoted Text preference	+2.20
Recency Preference	+2.15
Nominative Case preference	+1.85
First NP preference	+1.40
Nominal predicate preference	+1.20
Repetition preference	+1.20 if repletion is more than once
Punctuation preference	+1.15
Antecedent of Zero pronoun preference	+1.05

The following four steps or modules are used in this work to resolve pronouns by taking preprocessed Turkish texts as input:

i. Sentence splitting
ii. Pronoun extraction i.e. third person personal pronouns and reflexive pronouns
iii. Creating candidate list of antecedents
iv. Determining antecedents

The performance of the system is tested on two different sample texts. On the first test sample, the system correctly resolved 162 out of 190 pronouns as success rate (Recall) which is 85.2 % and having 88% precision. On the second test sample, it correctly resolved 151 out of 205 pronouns as success rate (Recall) which is 73.6 % and having 90.9 % precision.

3.2. Knowledge-poor Anaphora Resolution System for Estonian

Pilleriin Mutso [2] used knowledge poor anaphora resolution approach to resolve Estonian pronouns. The work focused on third person pronouns. The researcher used knowledge poor pronominal anaphora used in technical manuals for English as a baseline for this work with minor modifications. He used the terminology filters and indicators for constraint and preference rules. Filters represent constraint rules whereas indicators represent preference rules. Since gender constraint is not applied for Estonian language the only constraint rule used was number

agreement constraint rule. New preferences which are not found on the research done in English, baseline for this work, are added to the constraint and preference rules list of this work while removing some of the preference rules out because of the language difference. Preference rules used in this work are listed as follows.

- Frequency indicator (lexical reiteration)
- Pattern indicator
- Referential distance indicator
- Indicative verbs indicator
- Boost pronoun indicator
- Name indicator
- Declination indicator
- Quotes indicator

The following are list of preference rules not included in this work due to different reasons:

- Givenness indicator
- Syntactic parallelism indicator
- Section heading preference
- Term preference indicator
- Non prepositional noun phrase indicator
- Immediate reference indicator

Preference rules added to the system as new are Name indicator and quotes indicator
The corpus used in this work is manually annotated document. It is annotated morphologically and syntactically. The steps or modules used in this work are presented as follows.

i. Parsing the document using parsers
ii. Storing word and sentence objects
iii. Locating anaphora
iv. Locating candidates
v. Applying number agreement constraint rule
vi. Applying preference rules
vii. Applying recency coefficient and then sort by the scores

viii. Proposing the correct antecedent

To evaluate performance of the system the researcher used training and testing datasets. Dataset having 646 third person personal pronouns is used for training purpose whereas 856 third person personal pronouns is used for testing purpose. In training phase, out of 646 third person personal pronouns, the system resolved 440 third person personal pronouns correctly which is around 73.6%. In testing phase, out of 856 third person personal pronouns, the system resolved 549 third person personal pronouns correctly which is around 73.7%.

3.3. Automatic Pronominal Anaphora Resolution in English Texts

Tyne Liang and Dian-song Wu [7] used techniques like WordNet ontology to increase accuracy of anaphora resolution. In this work animate entities in the dataset are identified using WordNet ontology. It is also used to identify gender information of words. The proposed system deals with intra-sentential and inter-sentential anaphora in texts and also with pleonastic pronouns (it). Researchers used the terminology called heuristic rules to represent constraint and preference rules. The architecture designed by the researchers is presented on Figure 3.1. Constraint and preference rules used in this work are listed as follows.

- Syntactic parallelism rule
- Semantic parallelism rule
- Definiteness rule
- Mention frequency rule
- Sentence recency rule
- Non prepositional noun phrase rule
- Conjunction constraint rule

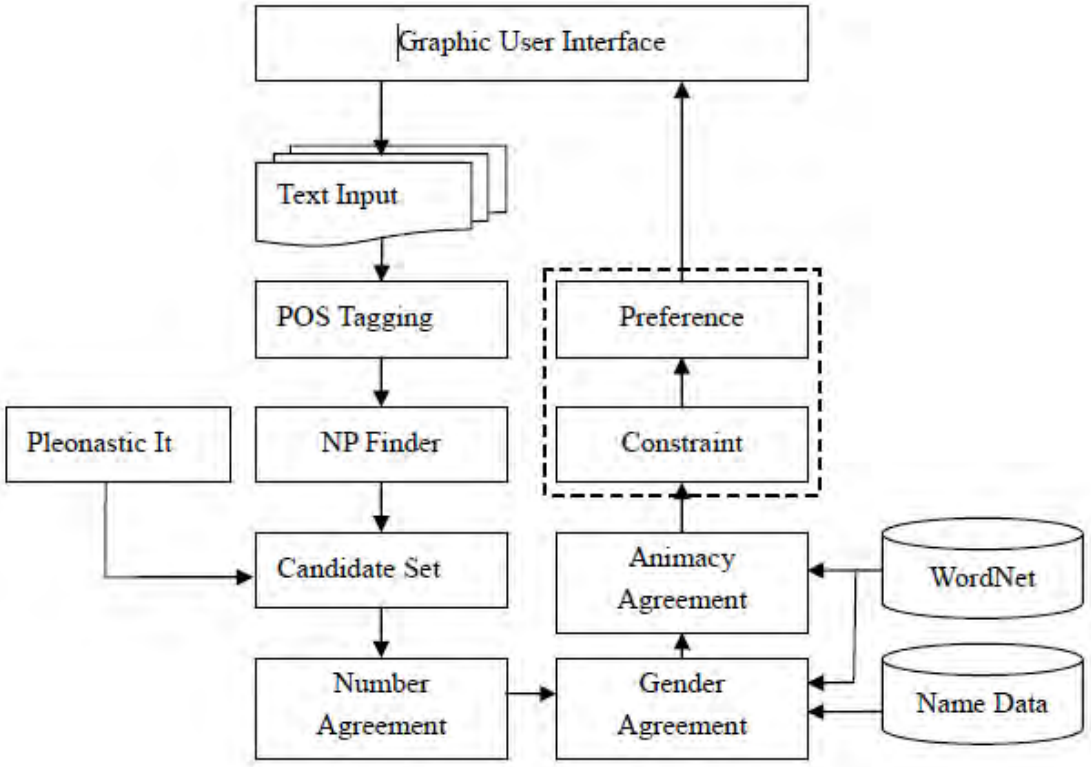

Fig. 3.1 System architecture for automatic pronominal AR in English texts [7]

The dataset used in this work includes texts from different genres having total of 14,124 words, 2,970 noun phrases and 530 anaphors. Using the proposed architecture presented in Fig.3.1, they achieved overall success rate of 77%.

3.4. Robust pronoun resolution with limited knowledge

Ruslan Mitkov [12] used knowledge poor anaphora resolution approach for pronoun resolution in English language. As it is presented in this work, the approach used here for English is applied to other languages successfully. The researcher discovered knowledge poor anaphora resolution approach for the first time to avoid labour intensive, expensive and time consuming tasks to resolve anaphors. Since the performance of the system using this approach in this work is good, it can be an evidence of how anaphors in a specific genre can be resolved quite successfully without any sophisticated linguistic knowledge.

As presented in the above paragraph, robust knowledge poor approach is developed to avoid complex syntactic, semantic and discourse analysis needed for anaphora resolution. This

approach starts with the output of part of speech tagger as input. After taking part of speech tagged texts, it searches for noun phrases which precede the anaphor within 2 sentences limit. Noun phrases identified within the scope are taken as candidate antecedents on which gender and number agreement constraints are applied. Finally, preference rules, named as antecedent indicators in this work, applied on candidate antecedents passed gender and number agreement rule, and noun phrase with highest aggregate score is proposed as the required antecedent. Preference rules used in this work are presented as follows.

- Definiteness
- Givenness
- Indicating verbs
- Lexical reiteration
- Section heading preference
- Non prepositional noun phrases
- Collocation pattern preference
- Immediate reference
- Referential distance
- Term preference

Informally described algorithm used in this work to solve pronouns is presented as follows:

1. Examine the current sentence and the two preceding sentences (if available). Look for noun phrases only to the left of the anaphor
2. Select from the noun phrases identified only those which agree in gender and number with the pronominal anaphor and group them as a set of potential candidates
3. Apply the antecedent indicators to each potential candidate and assign scores; the candidate with the highest aggregate score is proposed as antecedent. If two candidates have an equal score, the candidate with the higher score for immediate reference is proposed as antecedent. If immediate reference doesn‟t hold, propose the candidate with higher score for collocational pattern. If collocational pattern suggests a tie or does not hold, select the candidate with higher score for indicating verbs. If this indicator doesn‟t hold again, go for the most recent candidate.

The performance of this approach resulted in success rate of 89.7%. It is also applied to other languages like Polish and Arabic, and resulted in 93.3% and 95.2% success rates respectively.

3.5. Summary

As it presented in the works above, knowledge poor anaphora resolution approach is adopted to different languages effectively. The researchers used constraint and preference rules and other techniques like WordNet ontology that helped them solve anaphora resolution problem. Some of the constraint rules they used were language depedent while others were language independent. There are constraint and preference rules which can be applied to all languages whereas there are some that can't be applied to all languages.

To the best knowledge of the researcher so far no research is done in the area of anaphora resolution for Amharic language. Amharic language is morphologically complex and different than other languages like English. So, in this work we will employ knowledge poor anaphora resolution approach to design Amharic anaphora resolution model.

CHAPTER FOUR

AMHARIC ANAPHORA RESOLUTION MODEL

4.1. Introduction

As discussed in the previous chapter, various models have been proposed to resolve anaphora for different languages. The models depend on the characteristics of the languages. As a result, it is difficult to directly apply the proposed models to Amharic. Thus, we have proposed a new anaphora resolution model for Amharic. This chapter discusses about Amharic Anaphora Resolution (AAR) model and prototype development. The first part of this chapter discusses about architecture of AAR model. The second part of this chapter discusses about AAR prototype; how it is developed, the algorithm used and other related issues.

4.2. AAR architecture

As it is discussed on Chapter 2, Amharic personal pronouns can be found hidden inside words or independent. Hidden anaphors are bound morphemes that resulted in complexity of Amharic words. As a result, we are required to deal with both of the personal pronouns. Proposed Amharic AR architecture has two parts: resolution of independent anaphors and resolution of hidden anaphors. They show components used to resolve independent Amharic personal pronouns (independent anaphors) and hidden Amharic personal pronouns (hidden anaphors). The main components of the architecture are data preprocessing, identification of hidden anaphors, identification of independent anaphors, identification of candidate antecedents and Anaphora resolution (see Figure 4.1). The architecture shows the flow of components and subcomponents starting from input to output of the Amharic anaphora resolution process. Input Amharic texts to the model are preprocessed before they are used in the next levels of the componenets. Hidden and independent Amharic anaphors, and candidate antecedents for the anaphors are identified from the preprocessed dataset. The resolution process of anaphors is performed on the identified list of anaphors and candidate antecedents. Descriptions of all AR components are presented from Sections 4.2.1 to 4.2.3.

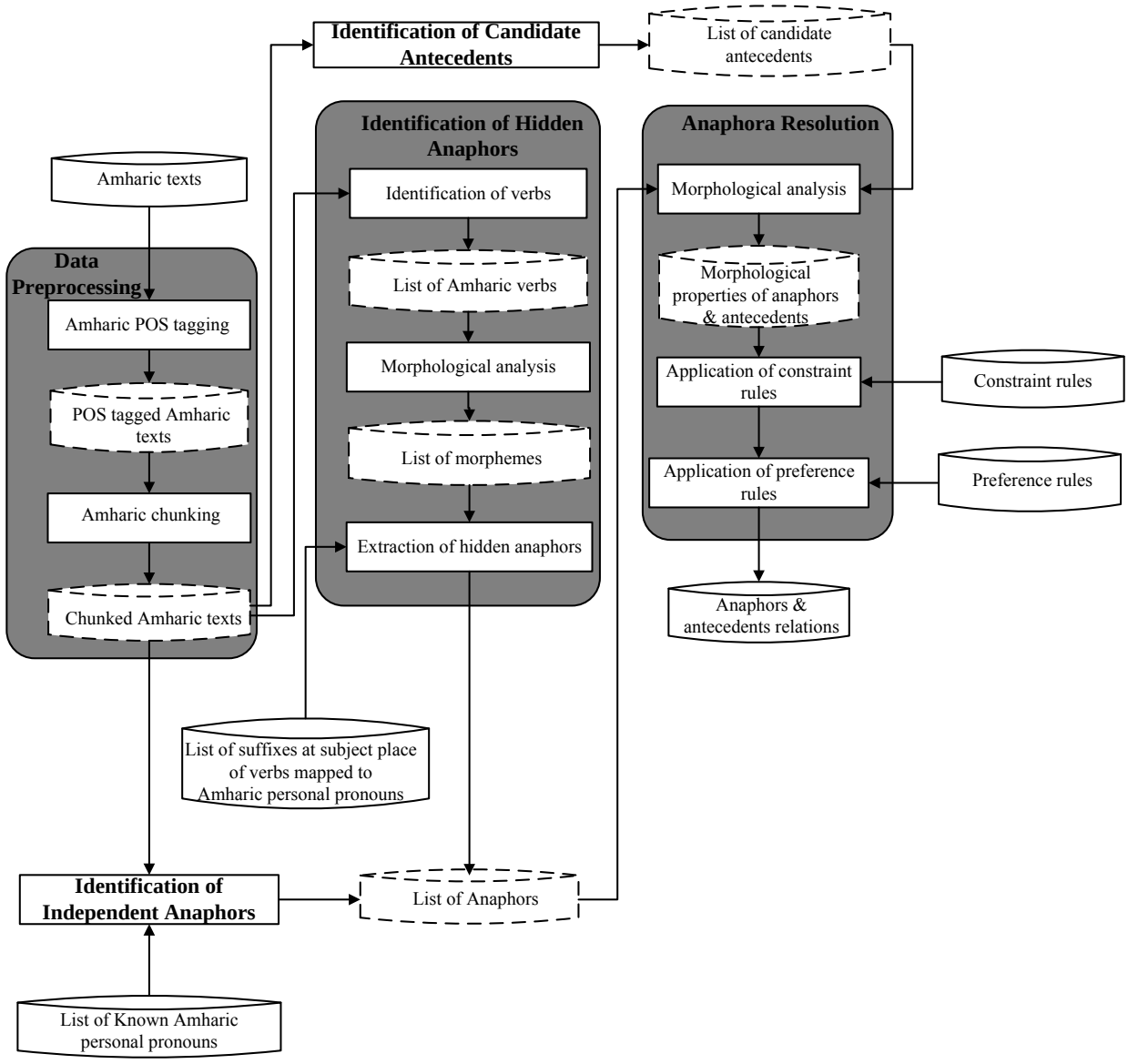

Fig. 4.1 Architecture of the Amharic Anaphora Resolution Model

4.2.1. Data preprocessing

Identification of candidate antecedents for the anaphors to be resolved needs the dataset to be preprocessed to get noun phrases and nouns which are considered as possible candidate

antecedents in this work. Tagging and chunking are tasks used to give different types of word classes and phrases like nouns and noun phrases respectively. Moreover, these tasks are used to identify verbs to extract hidden personal pronouns inside them. As a result, tagging and chunking are tasks used to preprocess Amharic input texts. Subcomponents of data preprocessing module presented on Fig. 4.1 of Amharic AR architecture are discussed in this section.

Amharic POS tagging

Part-of-speech (POS) tagging is a classification task with the goal to assign word classes to the words in a text [34]. We used POS tagging because it helps identify antecedents and independent anaphors, and used as input for chunking. Because of the unavailability of working POS tagger for Amharic, we tagged Amharic sentences manually.

Amharic chunking

Chunking is a Natural Language Processing (NLP) task that consists of dividing a text into syntactically correlated parts of words which are non-overlapping [19]. Chunking is needed to get noun phrases because identification of them helps identify candidate antecedents for the anaphors. For the identification and resolution of hidden anaphors, we used the ouput of Amharic chunker developed by Abeba Ibrahim and Yaregal Assabie [19]. However, we chunked Amharic sentences manually for the resolution of independent anaphors. It is because the output of Amharic chunker has no independent anaphors. The examples below show comparison made between chunked and non chunked Amharic sentences to show how chunking helps identify anaphors and antecedents.

Example:

- መስተዳድሩ ከፍተኛ ውጤት ላስመዘገቡ ተማሪዎች ሽልማት ሰጠ

  [(' መስተዳድሩ ' , ' N ') , (' ከፍተኛ ADJ ውጤት N ' , ' NP '), (' ላስመዘገቡ ' , ' VPREP ') , (' ተማሪዎች ', ' N ') , (' ሽልማት N ሰጠ V ' , ' VP ')]

- ሄሮድስ ጠቢባኑን አስጠርቶ ኮከቡ የታየበትን ትክክለኛ ጊዜ ተረዳ :: እኔም እንድሰግድለት የት እንዳለ ንገሩኝ አላቸው ::

  [(' ሄሮድስ ' , N) , (' ጠቢባኑን ' N ' አስጠርቶ ' , V ' , ' NP ') , (' ኮከቡ ' N 'የታየበትን ' VP ,' VP ') , (' ትክክለኛ ' ADJ ' ጊዜ ' N ' ተረዳ ' V , ' VP ')] ::

[(' እኔም ' PRO ' እንድሰግድለት ' VP ') (' የት እንዳለ ' VP ' ንገሩኝ ' V ' አላቸው' VP , VP)]
::

As it is discussed above, antecedents used in this work are nouns and head nouns of noun phrases; independent anaphors are Amharic personal pronouns whereas hidden anaphors are extracted from verbs. As a result, it is required to identify nouns to get antecedents whereas verbs and Amharic personal pronouns to get anaphors. Chunked Amharic sentences presented above show nouns, pronouns and verbs clearly but non chunked Amharic sentences don‟t. So, chunking helps identify antecedents and anaphors. More examples on the chunking ouput are presented below.

- [(' መስተዳድሩ ' , ' N ') , (' ከፍተኛ ADJ ውጤት N ' , ' NP '), (' ላስመዘገቡ ' , ' VPREP ') , (' ተማሪዎች ', ' N ') , (' ሸልማት N ሰጠ V ' , ' VP ')]
- [(' ሰራተኛዋ ' , ' N ') , (' ዝናብ N ያረጠበውን VREL ' , ' S '), (' ልብስ ' , ' N ') , (' ቤት N ውስጥ PREP ' , ' PP '), (' አሰጣች ' , ' V ')]
- [(' ልጅቷ ', ' N ') , (' ትልቅ ADJ ወንድሟ N ', ' NP ') , (' እንዲያስጠናት ' , ' V ') , (' ክፍል N ውስጥ PREP ' , ' PP ') , (' ነገረችው ' , ' V ')]
- [(' ሸምግሌው ' , ' N ') , (' ጥርሳቸው N ስላለቀ VPREP ' , ' VP ') , (' ብዙ ADJ ጥሬ N ' , ' NP ') , (' መብላት N አቆሙ V ' , ' VP ')]
- [(' ሄሮድስ ' , N) , (' ጠቢባኑን ' N ' አስጠርቶ ' , V ' , ' NP ') , (' ኮከቡ ' N 'የታየበትን ' VP ,' VP ') , (' ትክክለኛ ' ADJ ' ጊዜ ' N ' ተረዳ ' V , ' VP ')] ::

  [(' እኔም ' PRO ' እንድሰግድለት ' VP ') (' የት እንዳለ ' VP ' ንገሩኝ ' V ' አላቸው' VP , VP)]
  ::
- [(' ኢየሱስ ' , N) , (' በዮሐንስ ' NP ' ለመጠመቅ ' VP , ' VP ') , (' ከገሊላ ' NP ' ' ወደ ' PRE ዮርዳሮስ ' N , ' NP ') (' መጣ ' V)] ::

  [(' ዮሐንስ ' N ' ግን ' , ' NP ') , (' እኔ ' PRO ' ባንተ ' PRO ' መጠመቅ ' V ' ሲያስፈልገኝ ' V , ' VP ') (' እንዴት ' ' አንተ ' PRO ' ወደ ' PRE ' እኔ ' PRO ' ትመጣለህ ' V , ' VP ') , (' ብሎ ተከላከለ ' VP)] ::

4.2.2. Identification of independent anaphors

Independent Amharic personal pronouns are considered as independent anaphors. As discussed on chapter 2, independent Amharic personal pronouns are 8 in number without the personal pronouns indicating politeness. To perform the resolution of independent anaphors, identification of independent Amharic personal pronouns is needed. However, it is to be noted that Amharic personal pronouns are subject to various affixation processes. For example, many words can be derived through affixation processes from the personal pronoun እርሷ (*IrsWa*), some of which are: የእርሷ (*yeIrsWa*), ከእርሷ (*keIrsWa*), በእርሷ (*beIrsWa*), እርሷን (*IrsWan*), የእርሷን (*yeIrsWan*), እርሷንም (*IrsWanm*), የእርሷንም (*yeIrsWa*) , እንደእርሷ (*IndeIrsWa*), etc. Thus, identification of independent anaphors is achieved by having a database of Amharic personal pronouns (in their root forms) and matching them against the sub-patterns of personal pronouns in the given POS tagged text. Identified independent anaphors are stored in a temporary storage which will be used later in the process of anaphora resolution.

4.2.3. Identification of hidden anaphors

The main contribution of this work is resolving hidden anaphors. Identification of verbs and then extracting their morphological properties is needed to extract hidden anaphors inside verbs. The subcomponents of this module are identification of verbs, morphological analysis and extracting hidden anaphors. The subcomponents are discussed in this section.

Identification of verbs

As discussed on Chapter 2, Amharic verbs carry different kind of information on their affixes like person, number, gender, subject and others including personal pronouns. So, identification of verbs is needed to extract personal pronouns hidden inside them. As a result, the chunk result previously generated in the Data Processing component is used to identify verbs.

Morphological Analysis

After verbs are identified, their morphology is analysed so that hidden anaphors are extracted. Hornmorpho was used [6] to analyze the morphology of Amharic verbs. Examples listed below show morpheme information of verbs generated.

- ሰጠ: {sT*+1e2_e} (prf) - e (sb=3sm)

- ሰጣቸው : {sT*+1e2_e} (prf) - e (sb=3sm) -ac_ew (ob=3p)
- ልኮላታል : {l'k+13} (ger) - o (sb=3sm) -l _ (prep2) - at (ob=3sf) - al (aux)
- ጠፉ : {Tf'+1e2_a} (prf) - u (sb=3p)
- ነገረችው : {ngr+1e2_e3} (prf) - ec _ (sb=3sf) - w (def|ob=3sm)
- አቆሙ : {qwm+a1o3} (prf,trans) - u (sb=3p)
- ያከብራል : y (sb=3sm|3p) - {kbr+a1e23} (imprf,trans) -al (aux)

Extraction of hidden anaphors

As discussed under „identification of verbs" section, hidden anaphors are extracted from verbs. So, this component identifes anaphors which are personal pronouns hidden inside verbs. The suffix information of verbs is used to identify them. The following example shows how hidden anaphors are extracted.

Example,

ሰጠ: {sT*+1e2_e} (prf) - e (sb=3sm)

In the above example, „e" is found at subject place (**sb**=3sm) of the word ሰጠ. If suffix „e" is found on subject place of the word like the above example it refers to the personal pronoun „እሱ (He)". As a result, personal pronoun or anaphor hidden inside the verb ሰጠ is „እሱ (He)"

List of possible suffix information used to extract hidden anaphors is presented on Table 4.3. See Section 4.3 for detail explanation on how hidden personal pronouns are extracted.

4.2.4. Identification of candidate antecedents

Identification of candidate antecedents is preparation of lists so that the correct antecedents referred by anaphors are selected from it. All nouns and head nouns of noun phrases are used as possible candidate antecedents. So, this component is used to find possible antecedents which could be best antecedent of the anaphor to be resolved. The following example shows antecedents of the sentence.

Example,

[(' መስተዳድሩ ' , ' N ') , (' ከፍተኛ ADJ ውጤት N ' , ' NP ') , (' ላስመዘገቡ ' , ' VPREP ') , (' ተማሪዎች ' , ' N ') , (' ሸልማት N ሰጠ V ' , ' VP ')]

In the example above, መስተዳድሩ, ውጤት, ተማሪዎች and ሸልማት are nouns and head nouns of noun phrases identified. As a result, they are candidate antecedents for the anaphor hidden inside the verb „ሰጠ".

4.2.5. Anaphora resolution

As discussed on chapter 1, anaphora resolution is the task of resolving anaphors to their correct antecedents. So, this component resolves anaphors to their correct antecedents based on identified candidate antecedents and anaphors. The subcomponents of this component are applicaton of preference rules, application of constraint rules and morphological analysis.

Morphological analysis

Gender, number, definiteness and person information of words are bound morphemes in Amharic. To identify those information, morphological analysis is needed. So, this subcomponent is used to get gender, number, definiteness and person information of anaphors and antecedents. The following examples show gender, number, definiteness and person information extracted.

Example 1,

Word: ሰጠ
POS: verb, root: <sT*>, citation: ሰጠ
Subject: 3, sing, masc
Grammar: perfective
The above output is anal (anal is short for analyze which is a function used in Hornmorpho to analyze a word) output that Hornmorpho produced. In the example gender, number and person information of the word ሰጠ are clearly presented; it is third person singular and masculine gender.

It is also possible to get the gender, number and person information from the seg (seg is short for segmentation which is function used in Hornmorpho to segment a word in to small meaning full pieces) output of Hornmorpho as presented below.

Example 2,

ሰጠ: {sT*+1e2_e} (prf) - e (sb=3sm)

In the above ouput (sb=3sm) shows gender, number and person information of the given word.

It is also possible to get definite information of nouns from the anal output of Hornmorpho. The following example shows the anal output of Hornmorpho.

Example 3,

Word: ሰራተኞቹ

POS: noun, stem: ሰራተኛ

Grammar: plural, definite

As it is shown above, the noun ሰራተኞቹ is definite.

After looking at the output of Hornmorpho to get these information, we decided to tag manually because Hornmorpho is not effective in giving those information.

Application of constraint rules

Constraint rules are used to discard antecedents that don‟t agree with the anaphors. They are applied both on anaphors and candidate antecedents to select the correct antecedents or antecedents that satisfy constraint rules. The antecedent-anaphor relationship is stored in database as correct relationship if only one antecedent is left after the constraint rules are applied. The constraint rules used in this work are gender, number and person agreements. They are presented as follows:

1. Anaphors and antecedents should agree in gender information
2. Anaphors and antecedents should agree in number information
3. Anaphors and antecedents should agree in person information.

Application of preference rules

As discussed on chapter 3, preference rules are rules that give more preference to antecedents that satisfy the rule while give nothing or negative value that don‟t satisfy the the rule. This subcomponent is used to select correct antecedent referred by the anaphor by applying preference rules when more than one antecedent‟s are left as a choice after constraint rules are applied. The antecedent-anaphor relationship is stored in database as correct relationship after the preference rules are applied. Preference rules used in this work are definiteness, subject place, recency, boost pronoun and mention frequency. The preference rules are stored in a database with their respective values. The technique used to decide values of preference rules is discussed in chapter 5. Before discussing on the preference rules and their values of Amharic, we have discussed values of preference rules used in other languages.

In English language, if the antecedent is definite then 0 is assigned. If the antecedent is at subject place then a value of +1 is given to the antecedent. If an antecedent is more recent to the anaphor then +1 value is assigned to the antecedent [12].

Among the preference rules used in knowledge poor anaphora resolution system for Estonian, frequency indicator (lexical reiteration) has similar concept with mention frequency rule used in our work. For Estonian system the candidate was assigned a score of +1 when it is occurred in the section twice (section is considered to be a range from one heading to another). But, when the candidate occurred more than twice in the section, the candidate is given a score of +2 [2]. The same value is also used for English language [12]. Pronoun candidates (Boost pronoun preference rule) are assigned a score of +1 in Estonian language [2]. Preference rules used in our work are discussed as follows.

Subject place preference: in this work, the first noun or the first noun phrase in a sentence is considered as subject of the sentence. Noun at subject position of the sentence is more preferred than nouns not found at subject positions. If a noun is at subject place of the sentence then value of + 2.88 is assigned.

Definiteness: Indefiniteness is unmarked in Amharic, but definiteness is always marked by a suffix called the definite marker [38]. Unlike English there is no separate definite marker in

Amharic. Definite nouns are more preferred than nouns which are not definite. If a noun is definite then value of + 0.94 is assigned.

Recency: is a factor which checks whether the antecedent is more recent to the anaphor or not. The noun which is the most recent to the anaphor is given more weight than others which are not recent. If a noun is nearer to the anaphor then value of + 0.89 is assigned.

Boost pronoun: when pronouns are used as antecedent for the anaphors, then they are called as boost pronouns. This factor checks the occurrences of personal pronouns before the anaphors; the pronouns themselves are considered as possible candidates. One reason for that is the fact that pronouns indicate salience, as they represent noun phrases. The other reason is that one pronoun can refer to another pronoun that can refer to a noun phrase i.e. the anaphora can be solved transitively [2]. If the antecedent is boost pronoun then value of + 0.12 is assigned.

Mention frequency: is number of repetition of words. This preference rule checks whether the antecedent is found repetitively on the defined scope of sentences or not. The antecedent mentioned repetitively is given more priority. If the antecedent is mentioned repetitively (more than once) in a given scope then value of + 0.02 assigned.

4.3. Description of the AAR prototype

Designed Amharic AR prototype has two parts: resolution of hidden anaphors and resolution of independent anaphors. Algorithm presented on Fig. 4.6 which is informally described Mitkov"s robust, knowledge-poor algorithm [12] is used to solve both hidden as well as independent anaphors. It is modified to solve hidden Amharic anaphors. For the case of independent Amharic anaphors, the concept is almost used directly except few modifications in step 3.

1. Examine the current sentence and the two preceding sentences (if available). Look for noun phrases only to the left of the anaphor.
2. Select from the identified noun phrases only those which agree in gender and number with the pronominal anaphor and group them as a set of potential candidates.
3. Apply the antecedent indicators to each potential candidate and assign scores; propose the candidate with the highest aggregate. If two candidates have an equal score, propose the candidate with the higher score for immediate reference. If immediate reference does not hold, propose the candidate with higher score for collocational pattern. If collocational pattern suggests a tie or does not hold, select the candidate with higher score for indicating verbs. If this indicator does not hold again, go for the most recent candidate.

Fig 4.5 Mitkov‟s robust knowledge poor algorithm [12]

In the resolution of hidden anaphors part of the prototype, verbs are searched from sentences in the dataset to extract hidden anaphors inside them. The identified verbs are then given to morphological analyzer to get morphemes of a verb. From a set of extracted morpheme information, suffixes attached to verbs are used to get hidden anaphors. The algorithm used for the resolution of hidden anaphors is presented in Fig. 4.6. Examples presented below show how to extract hidden anaphors from verbs after morphemes of verbs are extracted.

Example,

1. ሰብለ እኔ ዩኒቨርስቲ እስከመጣሁበት ሰዓት ድረስ አጫወተችን

In the above sentence the verb is **አጫወተችን**. For this verb, Hormorpho gives the following morphemes.

አጫወተችን: {Cwt+a1a2_e3} (prf, trans) - ec _ (sb=3sf) - n (ob=1p);

As shown in the output, the morpheme „ec‟ is the suffix attached to the root word „Cwt‟ which is found in the subject place of a verb. It refers to the personal pronoun „እሷ„ This personal pronoun is second person singular and having Female gender. See Table 4.4 for list of all suffixes at subject place and mapping to their respective personal pronouns used in this work.

2. የትምህርት ሚኒስቴር ለኤችአይቪኤድስ መከላከያ 10ነጥብ8ሚሊየን ብር መደበ

In the above sentence the verb is መደበ. The morphemes of this verb that Hornmorpho generated are given as follows:

መደበ: {mdb+1e2_e3} (prf) - e (sb=3sm).

The morpheme found in subject place for this verb is „e‟ which is the suffix attached to the root word. This suffix refers to the personal pronoun ‚እሱ,„ It is second person singular and having male gender.

3. የእሱ በጎች ምሳ እንደበሉ ጠፉ

In the above sentence the verb is ጠፉ. The morphemes of this verb that Hornmorpho generate are given as follows:

ጠፉ: {Tf'+1e2_a} (prf) –u (sb=3p)

As it is shown in the Hornmorpho output above, the morpheme „u‟ is the suffix attached to the root word „Tf‟ which is found in the subject place. This suffix refers to the personal pronoun ‟እነሱ‟. It is second person plural.

4. እኛ አስተማሪው የሰጠንን የቤት ስራ ክፍል ውስጥ ሰራን

In the above sentence the verb is ሰራን. The morphemes of this verb that Hornmorpho generate are given as follows:

ሰራን፡ {sr'+1e2_a} (prf) –n (sb=1p)

As it is shown in the Hornmorpho output above, the morpheme „n‟ is the suffix attached to the root word „sf‟ which is found in the subject place of suffix order. This suffix refers to the personal pronoun ‚እኛ‟. It is first personal plural. In this case we can see as a personal pronoun can refer to another personal pronoun.

1. Identify verbs found on sentences and decompose into possible morphemes
2. Identify suffixes attached to the root word
3. Identify the suffix at the subject place of suffixation order
4. Compare suffixes identified with the predefined suffixes
5. Based on suffixes, identify hidden anaphors
6. Identify antecedents preceding anaphors within the defined scope
7. Apply constraint rules
8. If only one antecedent is left after the application of constraint rules, then take that antecedent as the correct antecedent referred by the anaphor
9. If there are more than one antecedentss left after the application of constrain rules, apply preference rules on those antecedents
10. If there is only one antecedent having highest preference rule value after the application of preference rules then that antecedent is selected as correct antecedent. If two or more than two candidate antecedents have same highest aggregate preference rule value, then the antecedent in subject place of the sentence is selected as the correct antecedent referred by the anaphor. If it doesn't hold, then antecedent which is definite noun is selected as correct antecedent. If definite noun is also doesn't hold, antecedent which is more recent to the anaphor is selected. If all these don't hold, the antecedent which is boost pronoun is selected. Again if all the above cases don't hold, the antecedent which is mentioned frequently is selected. Finally, if all the above criteria's doesn't hold then it is concluded as anaphora can't be resolved

Fig 4.6 Resolution of hidden Amharic anaphor‟s algorithm

5. ሰብለ እኔ ዩኒቨርስቲ እስከመጣሁበት ሰዓት ድረስ አጫወተችን

In the above sentence እስከመጣሁበት is the verb. The morphemes of this verb that Hornmorpho generate are given as follows:

እስከመጣሁበት: 'ske (prep1)-(rel)-{mT'+1e2_a} (prf)-hu (sb=1s)-b_(prep2)-et (ob=3sm)

As it is shown in the Hornmorpho output above, the morpheme „hu‟‟is the suffix attached to the root word „mT‟‟ which is found in the subject place of suffix order. This suffix refers to the personal pronoun „እኔ‟‟. It is first person singular.

So, all the personal pronouns or anaphors hidden inside verbs are extracted as the examples provided above.

After the hidden personal pronouns are known, the next step is identifying candidate antecedents for the extracted anaphors. To identify the antecedent that is being referred by hidden anaphor, all nouns and head nouns of noun phrases preceding the anaphor are identified.

When anaphors and candidate antecedents are identified, the next step is applying constraint and preference rules to resolve anaphors. The resolution process depends on the constraint and preference rules according to the knowledge poor anaphora resolution approach. As it is discussed in the previous chapters, knowledge-poor anaphora resolution approach is the type of anaphora resolution approach that doesn‟‟t use much of linguistic and semantic knowledge to solve anaphors. Constraint rules are applied on identified anaphors and candidate antecedents to discard the antecedents that don‟‟t match with the anaphors. Constraint rules which are gender, number and person agreement are applied on set of nouns selected as possible candidates. Nouns matching with hidden anaphors in gender, number and person passes to the next step of the resolution where the preference rules applied.

After the application of constraint rules, if there is only one antecedent left then that antecedent is considered as the best antecedent being referred by the anaphor. But, if there are more than one antecedents left after the application of constraint rules, then preference rules are applied on those antecedents. As discussed on section 4.2, preference rules selected for this work are subject place preference, definiteness, recency, boost pronoun and mention frequency. The attached values of these preference rules are applied on antecedents; they may give more preference to the antecedent by adding a positive value if it satisfies the criteria or give no value if it doesn‟‟t satisfy the criteria. The antecedent having largest sum of preference values is selected as the best or correct antecedent referred by the anaphor. When all the candidate antecedents have same largest preference values, then antecedent which is found on subject place of the sentence is

considered as best antecedent of the anaphor; if there is no antecedent at subject place of the sentence then the antecedent which is definite noun is considered as correct antecedent; if the antecedent is not definite then more recent noun to the anaphor is considered as correct antecedent. The examples below show the resolution process of hidden anaphors.

Example,

1. መስተዳድሩ ከፍተኛ ውጤት ላስመዘገቡ ተማሪዎች ሽልማት ሰጠ

In the example above, the verb is ሰጠ where personal pronoun ‚እሱ" is hidden in. መስተዳድሩ, ውጤት, ተማሪዎች and ሽልማት are nouns and head nouns of noun phrases identified. Gender, number and person information of the verb ሰጠ is male, singular and third person pronoun respectively which is respective information of the hidden personal pronoun. Whereas the gender, number and person information of noun መስተዳድሩ is male, singular and third person, ውጤት is male, singular and third person, ተማሪዎች is plural and third person and ሽልማት is male, singular and third person. From the four set of candidate antecedents the noun ተማሪዎች is discarded when constraint rules are applied because the number information of it doesn"t match the respective information of the anaphor. The remaining three candidates pass to the next step where preference rule values are applied. The antecedent መስተዳድሩ is found at subject place of the sentence, moreover it is definite noun but not recent, boost pronoun or mentioned frequently. So, the aggregate value assigned to this candidate is 3.82 (definite= +0.94, subject place= +2.88, mention frequency=0, recent=0, boost pronoun=0). In the same manner, ውጤት is assigned 0, ሽልማት is assigned +0.89. From the three candidate antecedents, the candidate having largest value is መስተዳድሩ. As a result, it is selected as the antecedent being referred by the personal pronoun ‚እሱ" hidden inside verb ሰጠ.

2. የእሱ በጎች ምሳ እንደበሉ ጠፉ

In the above example, the verb is ጠፉ where personal pronoun ‚እነሱ,, is hidden in. በጎች and ምሳ are nouns and head nouns of noun phrases identified. Number and person information of the verb ጠፉ is plural and third person which is respective information of the hidden personal pronoun. Gender, number and person information of noun በጎች is plural and third person, and ምሳ is male, singular and third person. From the candidates only በጎች match or satisfy constraint rule; it matches with the gender, number and person information of verb or personal pronoun. Since

only one noun left after the application of candidate rules, the preference rules will not be applied here and the only noun left is selected as the antecedent referred by the personal pronoun ‚እነሱ„ hidden inside verb ጠፉ.

As it is shown in Table 4.3, there are different prefixes and suffixes attached to a verb [32]. There are subject markers both on prefixes and suffixes. In this work, we have focused only on suffixes of subject markers to identify personal pronouns hidden in verbs. Suffixes listed in Table 4.3 are not exhaustive.

In this work, we have added the suffixes „-hu‟ and „- c‟ as the suffix found on subject place of the verb. They are not found on the original work of Wondwossen Mulugeta and Michael Gasser. The following example shows how they are used.

Example,

- እኔ የክፍል ጓደኛዬ የተኛበትን አልጋ ከጎረቤት ተዋስኩ

In the above example ተዋስኩ is verb. Its morphemes are presented as follows:

ተዋስኩ : {w's+te1a3} (prf,pas) – hu (sb=1s).

The morpheme „-hu‟ is found in subject place of the verb ተዋስኩ which refers to first person singular.

- ቆንጆት ሰርጓ ስለደረሰ ከጓደኞቿ ጋር ሽርጉድ አለች

In the above example አለች is verb. Its morphemes are presented as follows:

አለች : {al_e} (pres) - c _ (sb=3sf)

The morpheme „- c‟ is found in subject place of the verb አለች which refers to third person singular.

Table 4.3 List of prefixes and suffixes of Amharic verbs [32]

Prep Conj	**Rel**	**Neg**	**Subj**	**Root Template Vocal**	**Subj**	**Appl**	**Obj**	**Neg**	**Conj**
be le ke ?nde	Ø ye yem ?m	Ø al ay at an sa ba	Ø te a as ? ti yi ?n Ø li ti yi ?n -a -as -a -as	$C_1C_2C_3$ 10101 1011 1101 ee e ea	E eh ex o a ew achu en ku k x e ec u achu n alehu aleh alex al aec alu alachu alen Ø i u	Ø l b	Ø N h x w t at achu acew ln wot	Ø m	Ø na n

In Table 4.4 below the suffixes in the subject position from Table 4.3 are mapped with their respective personal pronouns.

Table 4.4 Mapping of suffixes at subject place of verbs to personal pronouns

Suffixes	Example in Roman text	Example in Geez text	Personal pronouns matched
-E	belcE	በልቼ	እኔ
-eh	Belteh	በልተህ	አንተ
-ex	Beltex	በልተሽ	አንቺ
-o	Belto	በልቶ	እሱ
-a	Belta	በልታ	እሷ
-ew	Beltew	በልተው	እነሱ
-achu	Beltacu	በልታቹ	እናንተ
-en	Belten	በልተን	እኛ
-ku	„Safku	ፃፍኩ	እኔ
-k	„Safk	ፃፍክ	አንተ
-x	„Safx	ፃፍሽ	አንቺ
-e	„Safe	ፃፈ	እሱ
-ec	„Safec	ፃፈች	እሷ
-u	„Safu	ፃፉ	እነሱ
-achu	„Safacu	ፃፋቹ	እናንተ
-n	„Safn	ፃፍን	እኛ
-alehu	„Sfyalehu	ፅፍያለሁ	እኔ
-aleh	t‟‟Sfaleh	ትፅፋለህ	አንተ
-alex	t‟‟Sfyalex	ትፅፍያለሽ	አንቺ
-al	y‟‟Sfal	ይፅፋል	እሱ
-aec	t‟‟Sfalec	ትፅፋለች	እሷ
-alu	y‟‟Sfalu	ይፅፋሉ	እነሱ
-alachu	t‟‟Sfalachu	ትፅፋላቹ	እናንተ
-alen	Int‟‟Sfalen	እንፅፋለን	እኛ
-i	gdey(gdeli)	ግደይ	አንቺ

The resolution of independent anaphors or independent Amharic personal pronouns part of the prototype starts by identifying independent anapahors and candidate antecedents in the dataset.

All nouns and head nouns of noun phrases are used as candidate antecedetns. The constraint and preference rules used in the first part of the AR prototype are also used in this part. The algorithm used for the resolution of independent anaphors is presented on Fig. 4.7.

The following example shows the resolution of independent anaphors.
Example,

ነ. እግዚአብሔር የባረካቸው ምድርን ይወርሳሉ :: እርሱ የሚረግማቸው ግን ይጠፋሉ ::

In the above sentence, the independent anaphor is እርሱ (He). እግዚአብሔር and ምድርን are the candidate antecedents. The gender, number and person information of the anaphor እርሱ is third person singular masculine, whereas person, number and person information of the candidate antecedents እግዚአብሔር and ምድርን is third person singular masculine. Both candidates satisfy the constraint rules i.e. the gender, number and person information of anaphors matches the gender, number and person information of candidate antecedents. As a result, they pass to the next step where preference rules are applied to select the correct antecedent. The antecedent እግዚአብሔር is at subject place of the sentence but not definite, recent, boost pronoun or mentioned frequently so it is assigned an aggregate value of + 2.88 (definite= 0, subject place= +2.88, mention frequency=0, recent=0, boost pronoun=0). The candidate antecedent ምድርን is assigned an aggregate value of + 0.89 (definite= 0, subject place= 0, mention frequency=0, recent= + 0.89, boost pronoun=0). From the two candidates, the antecedent እግዚአብሔር has higher aggregate value than the antecedent ምድርን. As a result, the antecedent እግዚአብሔር is selected as correct antecedent referred by the anaphor እርሱ (He).

1. Identify independent Amharic anaphors found in sentences
2. Identify antecedents preceding the identified anaphors within the defined scope
3. Apply constraint rules
4. If only one antecedent is left after the application of constraint rules, then take that antecedent as the correct antecedent referred by the anaphor
5. If the antecedents left are more than one after the application of constraint rules, apply preference rules on those antecedents.
6. If there is only one antecedent having highest preference rule value after the application of preference rules then that antecedent is selected as correct antecedent. If two or more than two candidate antecedents have same highest aggregate preference rule value, then the antecedent in subject place of the sentence is selected as the correct antecedent referred by the anaphor. If it doesn't hold, then antecedent which is definite noun is selected. If definite noun is also doesn't hold, antecedent which is more recent to the anaphor is selected. If all these don't hold, the antecedent which is boost pronoun is selected. Again if all the above cases don't hold, the antecedent which is mentioned frequently is selected. Finally, if all the above criteria's doesn't hold then it is concluded as anaphora can't be resolved.

Fig 4.7 Resolution of independent Amharic anaphor"s algorithm

CHAPTER FIVE

EXPERIMENT

1.1. Introduction

In this chapter, we discuss about dataset gathering and preparation, technique used to prepare the training and testing dataset, tools used to implement the proposed model, evaluation techniques used to test the performance of the Amharic AR model and experimental results, and finally discuss about issues reduced performance of the model. We collected 311 sentences having 315 verbs tagged as verbs in the dataset that are used to extract hidden personal pronouns from verbs. From these sentences, 282 verbs known by Hornmorpho and having subject information on suffixes when extracted are used to perform the resolution process. Moreover, 163 sentences are collected to perform the resolution process of independent AR part. From these sentences, 110 personal pronouns are identified and used.

1.2. Dataset gathering and preparation

The dataset used in this work was collected from two sources: WIC and Amharic Holly Bible for training and testing purpose. The first set of dataset includes chunked Amharic sentences which were generated as output from Amharic chunker. A total of 320 sentences generated as output of Amharic chunker are used in this work. The developer of Amharic chunker collected sentences for the system from Amharic grammar books, research papers and WIC [19]. It includes sentences where there is no independent personal pronoun. This dataset is processed for the purpose of this work. The second set of dataset includes Amharic sentences gathered from Amharic holly bible. It includes sentences having independent personal pronouns. The reason for needing second classification of the dataset is because the output of Amharic chunker, which was collected from WIC and other sources mentioned above, doesn‟t have independent personal pronouns; there are very few independent personal pronouns in the output of Amharic chunker but they are not as needed for this work that is why we said there are no independent personal pronouns.

Sentences having independent personal pronouns are collected from Amharic holly bible. The collected sentences from the Amharic holly bible are manually tagged and chunked. A total of

163 sentences having 110 personal pronouns are collected from Amharic holly bible. They were collected from Amharic holly bible by the researcher, and tagged and chunked manually with the help of linguists. Gathering sentences and then, tagging and chunking was one of the challenging tasks during this research work.

The other set of dataset collected for the purpose of this work is the dataset used to decide values of preference rules for Amharic language. Since Amharic language is different from other languages like English, it was necessary to analyze and determine values of the preference rules used in this work.

As discussed on the previous chapter, the preference rules used in this work are definiteness, subject place, recency, boost pronoun and mention frequency. A total of 495 sentences are used to decide the values of preference rules. Out of the total sentences collected, 288 sentences having antecedents at subject place, 83 sentences having antecedents as recent nouns, 93 sentences having antecedents as definite nouns, 12 sentences having anaphors referring back to antecedents which are personal pronouns and 2 sentences where nouns mentioned more than once are selected as antecedents are identified. Each of the preference rule values are calculated manually using probability of distribution they have based on the collected sentences.

From the collected sentences, most of the personal pronouns refer to nouns at subject places. Definite nouns are the second largest set of class referred by personal pronouns. Recent nouns are ranked third to be referred by personal pronouns. The following table shows determined values of preference rules for Amharic AR.

Table 5.1 Values of preference rules for Amharic language

Preference Types	Total count in a dataset	Percent from 5
Subject place	288	+ 2.88
Definiteness	93	+ 0.94
Recency	83	+ 0.89
Mention frequency	2	+ 0.02
Boost Pronoun	12	+ 0.12

The following examples show some of the sentences used to decide the value of preference rules discussed above.

Personal pronouns referring nouns at subject place

1. ኮሚሽኑ የ1ነጥብ3ሚሊየን ብር የፕሮጀክት ስምምነት ተፈራረመ
2. አንባቢያን ደራሲው የሳለውን ጨካኝ ገጸ-ባህሪ ተአማኒነት ተጠራጠሩ
3. አስቴር የእህል ቤት ከፈተች
4. አስቴር እህቷ በጣም ስለወፈረች ከመጠን በላይ ተናደደች

As shown in the examples above, personal pronouns hidden inside verbs refer to the nouns at subject place. ተፈራረመ (he) refers to ኮሚሽኑ, ተጠራጠሩ (they) refers to አንባቢያን, ከፈተች (she) refers to አስቴር, ተናደደች (she) refers to አስቴር. For more examples see Appendix A

Personal Pronouns referring nouns which are definite

1. ድርጅቶቹ ከ57ሚሊየን ብር በላይ የፕሮጀክቶች ግንባታ አስጀመሩ
2. ገበያተኞቹ እኛ የቆምንበትን አውላላ ሜዳ ወድያው ሞሉት
3. ፖሊሱ ፈታሹ አደንዛዥ ዕፅ ያገኘበትን ግለሰብ ወሰደ
4. ወታደሮቹ ወደ ግቢያቸው ገቡ
5. መልእክተኞቹ ወደ ንጉሱ ተመለሱ :: እርሱ ግን ለምን እንደተመለሱ ጠየቃቸው ::
6. ዮሴፍ ወንድሞቹን ፍለጋ ሄደ :: ወንድሞቹም ዮሴፍ ወደ እነርሱ ሲመጣ በሩቅ አዩት ::

In the examples above, independent personal pronouns and also personal pronouns hidden inside verbs refer to nouns which are definite. As it is also shown in the examples, the subject place preference rule is highly related to the definite preference rules. Most of the nouns which are definite are also found on subject place. Definite noun ድርጅቶቹ is referred by አስጀመሩ(they) , ገበያተኞቹ is referred by ሞሉት(they) , ፖሊሱ is referred by ወሰደ(he) , ወታደሮቹ is referred by ገቡ(they). ንጉሱ is referred by እርሱ (he), ወንድሞቹም is referred by እነርሱ (they). For more examples see Appendix B

Personal pronouns referring to nouns which are recent

1. ካሳ የገዛው የትናንቱ በግ ታረደ
2. ክፍል ውስጥ የሚገኙት ጥቂት ተማሪዎች አሁን ወጡ
3. እነዚያ የእናንተ ጌሾዎች ተቆረጡ
4. ሚሳይል በሞቃዲሾ ፈንድቶ 14 ሰዎች ሞቱ
5. በዚያ ምሽት በአጋንንት የተያዙ ብዙ ሰዎችን ወደ ኢየሱስ አመጡ :: እርሱም መናፍስትን በቃሉ አስወጣ

In the examples above, independent personal pronouns and also personal pronouns hidden inside verbs refer to nouns which are most recent. ታረደ (he) refers to በግ, ወጡ (they) refers to ተማሪዎች, ተቆረጡ (they) refers to ጌሾዎች, ሞቱ (they) refers to ሰዎች, እርሱም (he) refers to ኢየሱስ. For more examples see Appendix C

Personal pronouns referring to nouns which are mentioned most (Mention frequency)

1. የሴፍ ወንድሞቹን ፍለጋ ሄደ :: ወንድሞቹም የሴፍ ወደ እነርሱ ሲመጣ በሩቅ አዩት ::
2. ኢየሱስ በዮሐንስ ለመጠመቅ ከገሊላ ወደ ዮርዳኖስ መጣ :: ዮሐንስ ግን እኔ ባንተ መጠመቅ ሲያስፈልገኝ እንዴት አንተ ወደ እኔ ትመጣለህ ብሎ ተከላከለ ::

In the examples above, independent personal pronouns refer to nouns which are mentioned more than once. The examples show the mention frequency preference rule. Nouns mentioned twice have more preference weight than nouns mentioned once. እነርሱ (they) refers to ወንድሞቹም, እኔ (I) refers to ዮሐንስ.

In this work the AR process is limited to one sentence for personal pronouns hidden inside verbs and two for independent personal pronouns. As a result, the dataset doesn‟t contain much of nouns mentioned more than once and selected as referred by anaphors. This preference rule is highly important if paragraph or more than one paragraph is given to AR system as input to be resolved.

Personal pronouns referring back to personal pronouns

1. እግዚአብሔር ሆይ ወደ አንተ ጮኽሁ :: አንተ ግን ዝም አልኸኝ ::
2. እኔ ካሳ የገዛውን ትንሽ በግ ትላንት አየሁት
3. እኔ የክፍል ጓደኛዬ የተኛበትን አልጋ ከጎረቤት ተዋስኩ
4. እኛ ስራ ሰራን

5. እሷ አባቷን ትመስላለች

In the examples above, independent personal pronouns and also personal pronouns hidden inside verbs refer back to independent personal pronouns. This preference rule is called as boost pronoun. አንተ (you) refers to አንተ, አየሁት (personal pronoun hidden is እኔ (I)) refers to እኔ, ተዋስኩ (personal pronoun hidden is እኔ (I)) refers to እኔ, ሰራን (personal pronoun hidden is እኛ (we)) refers to እኛ, ትመስላለች (personal pronoun hidden is እሷ (she)) refers to እሷ.

1.3. Implementation

To implement the proposed Amharic anaphora resolution model, we have used the following tools.

Python 3.2 programming language: is a programming language used to develop Amharic anaphora resolution prototype.

Amharic chunker: is a tool developed to chunk POS tagged Amharic texts [19]. It is used to get chunked amaharic texts. Sample outputs of the chunker are presented as follows.

- [(' ካሳ ' , ' N ') , [' በድንገት ' , ' NPREP '], (' ወደ PREP ጎጃም N ' , ' PP ') , (' ሄደ ' , ' V ')]
- [(' ፊታውራሪ ADJ አሰጌ N ' , ' NP ') , (' ሴት ' , ' N ') , (' በጣም ADJ ይወዳሉ V ' , ' VP ')]
- [(' በለጠ ' , ' N ') , (' ጭቃ N ላይ PREP ' , ' PP ') , (' ወደቀ ' , ' V ')]

Hornmorpho: is a tool developed using python programming language to analyze the morphology of three Ethiopian languages, Amharic, Oromo, and Tigrinya [6]. We have used this tool to get morpheme information of words. Sample morphemes extracted for the given words are presented as follows.

- ተቆረጡ : {qWrT+te1e2_e3}(prf,pas)-u(sb=3p)
- መጣች : {mT'+1e2_a}(prf)-ec_(sb=3sf)
- አየሁት : {'y*+a2_e}(prf) -hu(sb=1s)- t(def|ob=3sm)

SERA: is a tool developed to convert Ethiopic script into the seven bit American Standard for Computer Information Interchange. We have used it to convert Amharic texts into English texts. The following examples show Amharic texts and their respective converted texts.

- „ይጠፋሉ“ is converted as „yTefalu“
- „ተዋስኩ“ is converted as „tewasku“
- „አጫወተችን“ is converted as „aCawetecn“

1.4. Test results

A technique called 10-fold cross validation is used to prepare the dataset as training and testing. In 10-fold cross validation, the entire dataset is randomly partitioned into k equal samples. From these k samples, one sample is used as testing dataset whereas the remaining k-1 samples are used as training datasets. Since it is 10-fold cross validation, the process is repeated 10 times. Average value of the ten testing dataset results is used as overall performance of the system [22]. After preparation of the dataset for testing and training, the next step we performed was evaluating performance of the model. The evaluation methods we used to test performance of the Amharic AR model are Success Rate and Critical Success Rate [24, 25]. Their formula is presented as follows

$$\text{Success rate} = \frac{\text{Number of correctly resolved anaphors}}{\text{Number of all anaphors}} \quad (5.1)$$

$$\text{Critical success rate} = \frac{\text{Number of correctly resolved anaphors}}{\text{Number of anaphors with more than one antecedent}} \quad (5.2)$$

The second formula (5.2) is used when anaphora has more than one antecedent after the application of constraints; gender, number and person constraint rules. Critical success rate is less than or equals to the overall success rate. Anaphora resolution approaches having high critical success rate closer to the overall success rates can be taken as good approaches [24]. Since knowledge poor anaphora resolution approaches rely mostly on preference rules, it also shows how accurate the preference rules are in solving the antecedents to the correct anaphors.

While preparing dataset to evaluate performance of our Amharic AR model, hidden anaphora part, we have discarded some verbs which are not known by Hornmorpho and also verbs having no subject information in their suffixes. Some verbs like ናት, አለው, እያካሄደ, ነች, ተፈራረመ in the

dataset are not known by Hornmorpho while others like እንዲያስጠናት, ልትጠይቀው, ሳይበላ are known by Hornmorpho and morpheme information are extracted but their subject information is not in suffixes. Our scope is limited only to subject information of verbs in their suffixes. As a result, verbs having no subject information in suffixes are discarded. The above two cases cover around 10.5 % of the entire dataset. So we have worked on around 89.5% of the entire dataset. Summarized description is presented below in Table 5.2

Table 5.2 Accuracy of Hornmorpho to extract morpheme information of verbs

	Total count	Percent (%)
Verbs in dataset	315	100
Verbs having suffix information at subject place of verbs	282	89.52
Verbs having no subject information in suffix	10	3.17
Verbs not solved or known	23	7.30

The performance of Hornmorpho to extract morphemes from verbs based on the dataset provided by us is around 92.7%. As shown in Table 5.2 above, verbs‟ having no subject information in suffixes is around 3.17%. Out of these, some verbs have subject information in subject place of prefixes yet removed because it is out of our scope.

There is no separate personal pronoun to indicate the pleonastic pronoun „it‟ in Amharic. The pronoun he (እሱ) handles it. Unlike English, objects are referred by the personal pronoun „he‟ in Amharic. When personal pronouns hidden inside verbs are extracted, the personal pronouns referring objects are extracted as „he‟. So in the result of AR, objects are referred by personal pronoun „he‟. The same thing happens when resolving datasets having independent personal pronouns. The independent personal pronouns may refer to objects.

To test performance of our Amharic AR model, prototype is designed using python programming language. Its performance is evaluated using the evaluation methods or formulas

presented in Section 5.3. Both the hidden anaphora and independent anaphora resolution performances are presented in Sections 5.3.1 and 5.3.2.

1.4.1. Performance of Amharic AR model for anaphors hidden inside verbs

As discussed in introduction part (5.1) , 282 verbs are used to extract hidden personal pronouns inside them. The overall performance i.e. success rate of Amharic AR model for anaphors hidden inside verbs is 81.79 %. Generally, the success rate of Amharic AR model for anaphors hidden inside verbs is 81.79% whereas the critical success rate is 76.07% percent.

Amharic AR model correctly resolved 89 hidden anaphors from 117 hidden anphors having more than one antecedent after passing the constraint rules which is critical success rate. Detail performance result of Amharic AR model in the test datasets for hidden anaphors is presented on Table 5.3.

Table 5.3 Success rate result of Amharic AR model of hidden anaphors

Dataset	Wrongly resolved	Rejected by Amharic AR prototype	Success rate
Testing Dataset 1	4	1	82.14%
Testing Dataset 2	5	2	75%
Testing Dataset 3	4	0	85.71%
Testing Dataset 4	4	2	78.5%
Testing Dataset 5	7	2	67.86%
Testing Dataset 6	1	4	82.14%
Testing Dataset 7	2	1	89.29%
Testing Dataset 8	2	0	92.86%
Testing Dataset 9	1	1	92.86%
Testing Dataset 10	7	1	71.43%

1.4.2. Performance of Amharic AR model for independent anaphors

As discussed in introduction part (5.1), 110 independent personal pronouns are used as dataset for the independent Amharic AR part. The overall performance i.e. success rate of Amharic AR model for independent anaphors is 70.91%. Generally, the success rate of Amharic AR model for independent AR part is 70.91% whereas the critical success rate is 57.58%.

Amharic AR correctly resolved 46 independent anaphors from 70 independent anaphors having more than one antecedent after passing the constraint rules which is critical success rate. Detail performance result of Amharic AR model of independent AR part is presented on Table 5.4.

Table 5.4 Success rate result of Amharic AR model of independent anaphors

Dataset	Wrongly resolved	Rejected by Amharic AR prototype	Success rate
Testing Dataset 1	3	2	54.55%
Testing Dataset 2	1	0	90.91%
Testing Dataset 3	3	1	63.64%
Testing Dataset 4	2	1	72.73%
Testing Dataset 5	3	0	72.73%
Testing Dataset 6	6	0	45.45%
Testing Dataset 7	1	0	90.91%
Testing Dataset 8	4	0	63.64%
Testing Dataset 9	3	0	72.73%
Testing Dataset 10	2	0	81.82%

1.5. Discussion

During our experimental result analysis, we have found some reasons which led to the reduction in performance. The following points illustrate the reasons:

1. While working on the resolution of personal pronouns hidden inside verbs, we have found sentences talking about someone in singular way showing politeness. On these sentences verbs are represented as plural but semantically they are not plural rather they show politeness. Whereas, nouns that they are referring to are presented as singular. As a result, they don‟t match when constraint rules are applied. See the following examples.

   Example,

   ı. ሸምግሌው ጥርሳቸው ስላለቀ ብዙ ጥሬ መብላት አቆሙ

   As it is shown in the example above, personal pronoun hidden inside the verb አቆሙ is plural which is extracted as እነሱ (they) by our Amharic AR model. Whereas ሸምግሌው which should be resolved as correct antecedent is discarded because it is singular.

   In the above example, personal pronoun should have been extracted as እሳቸው (he) instead of እነሱ (they). But, Personal pronouns like እርስዎ (you) and እሳቸው (s/he) which show politeness are out of the scope of this work. As a result, they are not handled by our Amharic AR model which led to reduction in performance.

   2. ፕሬዝዳንቱ አምስት ሚኒስትሮችን በከፍተኛ ማዕረግ ሾሙ

   As it is shown in the above example, hidden personal pronoun inside the verb ሾሙ is identified as እነሱ (they) same to the first example. But, the candidate definite noun ፕሬዝዳንቱ which should be resolved as correct antecedent is discarded because it is singular when the anaphor is identified as plural.

2. There is scenario in Amharic where verbs and nouns not much in gender and number while sentences are correct. The number mismatch is discussed in the previous example. See the following example for gender mismatch.

   Example:

   1. አልቃይዳ በሶማሊያ እንደገና እንዳይደራጅ አሜሪካ ቅኝቷን አጠናክራለች

The above sentence is valid sentence. The verb አጠናከራለች has personal pronoun እሷ(she) hidden inside it. አሜሪካ is correct antecedent from existing choices preceding the verb and should also be selected as best antecedent by Amharic AR model. But Amharic AR model resolution process didn‟t select አሜሪካ as correct antecedent. It is because when the constraint rules, gender, number and person information, are applied it didn‟t match and as a result discarded. Gender, number and person information of አሜሪካ is 3rd person singular masculine. Gender, number and person information of the verb አጠናከራለች is 3rd person singular female.

3. There are sentences in our dataset which are valid semantically but antecedents being referred by the anaphors are not found in it. The following example elaborates more.
   1. ካሳ ለምን እንዳልመጣ አውቃለች

      In the above example, there is no antecedent being referred by the anaphor አውቃለች [እሷ (she)]. Personal pronoun hidden inside the verb አውቃለች is እሷ (she) but there is no antecedent in the sentence referred by it.

The performance, both success rate and critical success rate, presented in Section 5.3 are affected by the problems described above. The performance would have been better if limitations presented wasn‟t exist or solved. So, having all the limitations listed, the performance achieved during this work is encouraging.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

This section presents conclusion and recommendations of the work. The research work presented in this work is summarized briefly as conclusion, and the future works which would enhance performance of the model are presented as recommendations.

6.1. Conclusion

It was discussed that there are different forms of anaphora when dealing with anaphora resolution. But, in this work we have focused on only pronominal anaphora types of anaphora specifically on personal pronouns. To achieve our objective, we have used knowledge poor anaphora resolution approach. This approach doesn‟t need complex semantic, syntactic and other related knowledge‟s to solve anaphora. It uses low level knowledge‟s like morphological analysis.

The proposed Amharic anaphora resolution model has two parts. The first part is used to resolve independent Amharic personal pronouns whereas the second part is used to identify hidden Amharic personal pronouns and then solve them to their correct antecedents. The major components of designed anaphora resolution model, in this work, for Amharic language are data preprocessing, identification of hidden anaphors, identification of independent anaphors, identification of candidate antecedents and Anaphora resolution.

The model takes Amharic texts as inputs and then POS tags and chunks them to get chunked texts. Independent Amharic personal pronouns and hidden personal pronouns are identified from these chunked and POS tagged Amharic texts. Hidden Amharic personal pronouns are identified from verbs by extracting suffix information of verbs. To get hidden Amharic personal pronouns, we have applied Morphological analysis. Candidate antecedents are also identified from a set of chunked and POS tagged Amharic texts. All noun phrases and nouns in the POS tagged and chunked Amharic texts are considered as candidate antecedents for the anaphors. Constraint and preference rules are applied on a set of candidate antecedents and a set of anaphors to identify correct relationship between antecedents and anaphors.

To test the performance of the model, we have used the evaluation metrics called success rate and critical success rate. A technique called 10-fold cross validation is used to divide the dataset into training and testing datasets. 282 verbs are used for hidden anaphora resolution part whereas 163 independent personal pronouns are used for independent anaphora resolution part. The average success rate and critical success rate of hidden anaphora resolution part are 81.79% and 76.07% respectively whereas average success rate and critical success rate of independent personal pronouns resolution part are 70.91% and 57.58% respectively.

6.2. Recommendation

Anaphora resolution is very complex task which needs knowledge‟s like morphological, syntactic, semantic, world knowledge and other type of knowledge‟s which are complex to deal with. Applying all kind of knowledge‟s needed is labor intensive and time consuming, and as a result hard to achieve. As a result, we have focused on knowledge poor anaphora resolution approach. The following are some of the recommendations we suggest to be done in the future to increase the performance of the model.

- Performance of knowledge poor anaphora resolution approach depends highly on constraint and preference rules. Constraint and preference rules used in this work are very few in number. Therefore, we recommend incorporation of more constraint and preference rules, specially the rules which are specific to Amharic language.
- Amharic documents which are POS tagged and chunked are not found for Amharic language. As a result we tagged and chunked some documents manually which was so tiresome and time consuming. Therefore, we recommend incorporation of high performance POS tagging and chunking tools.
- Since we couldn‟t get Amharic documents which are POS tagged and chunked, and POS tagging and chunking texts manually is tiresome and time consuming, we have used dataset which is not large. We recommend preparation of large Amharic documents which are POS tagged and chunked so that it is used for anaphora resolution researches in the future.
- List of suffix information used to extract hidden anaphors from verbs presented in chapter 4 of this work are not exhaustive list. Adding more suffix rules found on subject

place of verbs helps capture more pronouns hidden inside verbs. Therefore, we recommend making suffix rules exhaustive as much as possible.

- In this work we focused on suffix information which are on subject place of suffixation order of verbs. Capturing prefix information which are found on subject place of verbs improves the performance of anaphora resolution. Therefore, we recommend using of prefixes on subject place of verbs combined with subject place information of suffixes.
- In this work we have extracted definiteness, gender, person and number information of nouns manually which was so tiresome and time consuming because we couldn‟t get high performance tool to do so. Here we recommend use of high performance tools to extract definiteness, gender, number and person information of nouns.
- In this work we have focused on only Amharic personal pronouns because trying to address all type of anaphora at one time is very complex and time consuming. We recommend extending the work to other types of anaphors.

References

[1] Elizabeth Duross Liddy (1990). "Anaphora in Natural Language Processing and Information Retrieval". *Information Processing and Management* 26(1): 39-52.

[2] Pilleriin Mutso (2008). "Knowledge-Poor Anaphora Resolution System for Estonian", Master‟s Thesis, University of Tartu, Tartu.

[3] Botley, Simon and Anthony Mark McEnery, editors (2000). Corpus-based and Computational Approaches to Discourse Anaphora. Amsterdam/Philadelphia: John Benjamins Publishing Company.

[4] Vaclav Nemcik (2006). "Anaphora Resolution", Master‟s Thesis, Masaryk University, Brno.

[5] Niklas Jakob and Iryana Gurevych (2010). "Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews", *In Proceedings of the ACL 2010 conference short papers*, Sweden: 263-268.

[6] Michael Gasser (2011). "HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya", *Conference on Human Language Technology for Development, Alexandria*, Egypt.

[7] Tyne Liang and Dian-Song Wu (2004). "Automatic Pronominal Anaphora Resolution in English Texts", *Computational Linguistics and Chinese Language Processing* (9)(1): 21-40.

[8] Jose L. Vicedo and Antonio Ferrandez (2000). "Importance of Pronominal Anaphora resolution in Question Answering systems", *In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong: 555-562

[9] Ruslan Mitkov, Richard Evans, and Constantin Orasan (2002). "A new, fully automatic version of Mitkov‟s knowledge-poor pronoun resolution method", *In Proceedings of the third International conference on Intelligent Text Processing and Computational Linguistics* , Mexico city, Mexico: 17-23.

[10] Dincheng Li, Tim Miller and William Schuler (2011). "A pronoun Anaphora Resolution System based on Factorial Hidden Markov Models", *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, Portland: 1169-1178.

[11] Niyu Ge, John Hale and Eugene Charniak (1998). "A Statistical Approach to Anaphora Resolution", *In Proceedings of the sixth workshop on very large copora:*161-170

[12] R.Mitkov (1998). "Robust Pronoun Resolution with limited knowledge", *In Proceedings of the 36th Annual Meeting of the Association for computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Canada: 869-875.

[13] W.Soon, H.Ng, and D.Lim (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases", *Computational Linguistics*, (27) (4):521-544

[14] Souha Mezghani Hammami, Lamia Hadrich Belguith, and Abdelmajid Ben Hamadou (2008). "Anaphora in Arabic Language: Developing a Corpora Annotating Tool for Anaphoric Links", *The international Arab conference on information technology.*

[15] Caroline V. Gaspering (2008). "Statistical Anaphora Resolution in biomedical texts". *In Proceedings of the 22nd International Conference on Computational Linguistics,* Stroudsburg, PA, USA: 257-264.

[16] Baldwin, Breck (1997). "CogNIAC : high precision coreference with limited knowledge and linguistic resources". *Proceedings of the ACL'97/EACL'97 workshop on Operational factors in practical, robust anaphora resolution*, Madrid, Spain: 38-45.

[17] Lappin, Shalom & Herbert Leass (1994). "An algorithm for pronominal anaphora resolution", *Computational Linguistics*: 535-561.

[18] Mitkov, Ruslan (1999), "Anaphora Resolution: The State of the Art," *Working paper (Based on the COLING'98/ACL'98 tutorial on anaphora resolution)*.

[19] Abeba Ibrahim and Yaregal Assabie (2013), "Hierarchical Amharic Base Phrase Chunking Using HMM with Error Pruning", *In Proceedings of the 6th Language & Technology Conference*, December 7-9, Poznan, Poland.

[20] Orăsan C (2007), "Pronominal anaphora resolution for text summarization", *In Proceedings of the recent advances on natural language processing:* 430–436

[21] Dilek Küçük (2005), "A knowledge-poor pronoun resolution system for Turkish", A thesis submitted to the Graduate School of Natural and Applied Sciences of Middle East Technical University.

[22] Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection" In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*), San Francisco, CA: Morgan Kaufmann: 1137–1143

[23] ባየ ይማ:: (2000 አ.ም) "የአማርኛ ሰዋሰው" የተሻሻለ ሁለተኛ እትም:: አዲስ አበባ እሌኒ ማተሚያ ቤት
Baye Yimam (2000), yeamarNa sewasew Eleni printing press, Addis Ababa, revised 2nd edition.

[24] Mitkov, R (2000), "Towards more comprehensive evaluation in anaphora resolution", *In Proceedings of The Second International Conference on Language Resources and Evaluation*, volume III, Athens, Greece: 1309-1314.

[25] Mitkov, Ruslan and Catalina Barbu (2001). "Evaluation tool for rule-based anaphora resolution methods", *In Proeedings of ACL'01*, Toulouse.

[26] Gordana Ilic Holen (2006), "Automatic Anaphora Resolution for Norwegian (ARN)", Thesis submitted for the degree of Candidata Philologiae, University of Oslo.

[27] Kabadjov, M.A. (2007) , "A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification", PhD thesis, Department of Computer Science, University of Essex.

[28] Nuno Nobre (2011), "Anaphora resolution", Master‟s thesis, Instituto Superior Tecnico / Universidade Tecnica de Lisboa.

[29] Tejaswini Deoskar. Techniques for Anaphora Resolution: A Survey, unpublished.

[30] Mitkov, Ruslan (2001). "Outstanding issues in anaphora resolution", *In Proceedings of the 2nd International Conference on Computational Linguistics and Intelligent Text Processing,* Springer, Berlin:110–125.

[31] H. Ho, K. Min, W Yeap (2004), "Pronominal Anaphora Resolution using a shallow meaning representation of sentences", *Proceedings of the 8th Pacific Rim International Conference on AI*: 862-871.

[32] Mulugeta, W and Gasser, M. (2012), "Learning morphological rules for Amharic verbs using inductive logic programming", *SALTMIL-AfLaT Workshop on Language Technology for Normalisation of Less-Resourced Languages*, Istanbul.

[33] Bjorn Gamback, Fredrik Olsson, Atelach Alemu Argaw, Lars Asker (2009), "Methods for Amharic Part-of-Speech Tagging", "Proceedings of the EACL Workshop on Language Technologies".

[34] Sisay Fissaha Adafre (2005), "Part of Speech Tagging for Amharic using Conditional Random Fields", *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages:* 47–54.

[35] Martha Yifiru Tachbelie (2010), "Morphology-Based Language Modeling for Amharic", Ph.D. thesis, University of Hamburg, Germany.

[36] Yitna Firdyiwek and Daniel Yacob (1997). "System for ethiopic representation in ascii".

[37] Atelach Alemu Argaw and Lars Asker (2007), "An Amharic stemmer: Reducing words to their citation forms", *Computational Approaches to Semitic Languages*, Prague, Czech Rep:104–110

[38] Ruth Kramer (2009). "Definite markers, phi-features, and agreement: a morphosyntactic investigation of the Amharic DP", Doctoral dissertation, UC Santa Cruz.

[39] Wondwossen Mulugeta, Gasser, M., and Baye Yimam (2012), "Incremental learning of affix segmentation", *International Conference on Computational Linguistics (COLING)*, Mumbai, India.

[40] Vaclav Nemcik (2007), "Anaphora Resolution for Czech", Dissertation Thesis, Masaryk University, Brno.

[41] G. Amare. "ዘ መና ዊ የ አ ማር ኛ ሰ ዋ ስ ው በ ቀ ላ ል አ ቀ ራ ረ ብ (Modern Amharic Grammar in a Simple Approach)", Addis Ababa, Ethiopia, 2010.

Appendix A: sample independent & hidden personal pronouns referring nouns at subject place

1. **መስተዳድሩ** ከፍተኛ ውጤት ላስመዘገቡ ተማሪዎች ሽልማት **ሰጠ**
2. **ሰብለ** እኔ ዩኒቨርስቲ እስከመጣሁበት ሰዓት ድረስ **አጫወተችን**
3. **ለማ** በሳል መምህር **ነው**
4. **በለጠ** ወደ ጎጃም **ሄደ**
5. **አስተማሪው** ለተማሪዎቹ መጽሀፍ **ሰጣቸው**
6. **አልማዝ** በጣም ደግ **ናት**
7. **በለጠ** ዛፍ ላይ **ወጣ**
8. **አትክልቱ** በደንብ **ለማ**
9. **ካሳ** ፈሪ **ነው**
10. **አበበ** ከወንድሙ ይበልጥ ሰው ፈሪ **ነው**
11. **አበራ** ወንድሙ የጋበዘውን ጎበዝ የኮሌጅ ተማሪ **ተዋወቀ**
12. **ቢሮው** ፈቃድ የተሰረዘባቸውን ነጋዴዎች ዝርዝር **አወጣ**
13. **ሰራተኛዋ** ዝናብ ያረጠበውን ልብስ ቤት ውስጥ **አሰጣች**
14. **ዘበኛው** አሰሪው ስለፈቀደለት የማታ ትምህርት **ተመዘገበ**
15. ትናንት የተገዛው ትልቅ የሀረር **ሰንጋ** ከመኪና እንደወረደ **ጠፋ**
16. **ሰውየው** እንደሄደ **ነገሩን**
17. **ካሳ** ትናንት በጎች **ገዛ**
18. ትናንት የመጣው **ልጅ** በባንክ ለአስቴር ገንዘብ **ልኮላታል**
19. የእሱ **በጎች** ምሳ እንደበሉ **ጠፉ**
20. **አስቴር** ጎጃም **ሄደች**
21. **ፕሬዚዳንቱ** አደጋው ህዝብ እንዳሳዘነ **ገለጹ**
22. **ትምህርትቤቶች** አገልግሎት መስጠት **ጀመሩ**
23. **ፋብሪካው** 85ነጥብ7ሚሊየን ብር የተጣራ ትርፍ **አገኘ**
24. **በለጠ** ትንሽ ቤት **አለው**
25. **በለጠ** ጭቃ ላይ **ወደቀ**
26. **አስተማሪው** ለተማሪዎቹ **ሰጣቸው**
27. **ካሳ** ቤት ውስጥ **ነበር**
28. **ሰውየው** በጣም ረጅም **ነው**
29. **ለማ** እንደ አባቱ ክፉኛ **ታመመ**

30. ፊታውራሪ **አሰጌ** ሴት በጣም **ይወዳሉ**
31. **አበበ** ሰው ፈሪ **ነው**
32. **ልጅቷ** ትልቅ ወንድሟ እንዲያስጠናት ከፍል ውስጥ **ነገረችው**
33. **አስቴር** እህቷ በጣም ስለወፈረች ከመጠን በላይ **ተናደደች**
34. **ወልዴ** ካሳ የያዘውን አዲስ ቦርሳ **አደነቀ**
35. **ሽምግሌው** ጥርሳቸው ስላለቀ ብዙ ጥሬ መብላት **አቆሙ**
36. **ቶሎሳ** አባቱ ጠንክሮ እንዳስተማረው ለአዲሷ ፍቅረኛው **ነገራት**
37. **ድርጅቱ** በ33ነጥብ5ሚሊየን ብር የመንገድ ግንባታዎችን **እያካሄደ** ነው
38. **ካሳ** አስቴርን **ገረፋት**
39. **ካሳ** በድንገት ወደ ጎጃም **ሄደ**
40. **አልማዝ** ከአስቴር የበለጠች ቆንጆ **ነች**
41. **ካሳ** እዚህ ቤት ውስጥ **ይኖራል**
42. **ካሳ** ቁርሱን ሳይበላ ወደ ትምህርትቤት **ሄደ**
43. **አህዴድ** ከጉባኤው ተጠናክሮ **ወጥቷል**
44. **ኢትዮጵያውያን** የጋራ መድረክ **መሰረቱ**
45. **እንግሊዝ** በኤርትራ የሚፈጸመውን እስራት **ተቃወመች**
46. ተዘዋዋሪ **ችሎቱ** 550 እስረኞች እንዲፈቱ **በየነ**
47. **ኮኮቦች** በንጋት በጣም **ያበራሉ**
48. **ሰውየው** ቤት ድረስ **ሸኘን**
49. **ካሳ** ሰው እጅግ **ያከብራል**
50. **ካሳ** ትንሽ መጽሀፍ **ሰጠ**
51. **ሄሮድስ** ጠቢባኑን አስጠርቶ ኮከቡ የታየበትን ትክክለኛ ጊዜ ተረዳ :: **እኔም** እንድሰግድለት የት እንዳለ ንገሩኝ አላቸው ::
52. **ኢየሱስ** በዮሐንስ ለመጠመቅ ከገሊላ ወደ ዮርዳሮስ መጣ:: ዮሐንስ ግን እኔ **ባንተ** መጠመቅ ሲያስፈልገኝ እንዴት አንተ ወደ እኔ ትመጣለህ ብሎ ተከላከለ::
53. **ኢየሱስ** አሥራሁለቱን ደቀመዛሙርት ወደ **እርሱ** ጠርቶ ርኩሳን መናፍስትን እንዲያወጡ ደዌንና ሕማምን ሁሉ እንዲፈውሱ ሥልጣንን ሰጣቸው::
54. **ሕዝቡም** በሙሉ ደሙ **በእኛና** በልጆቻችን ላይ ይሁን ብለው መለሱ::
55. **መልአኩም** ሴቶቹን እንዲህ አላቸው አትፍሩ የተሰቀለውን ኢየሱስን እንደምትፈልጉ አውቃለሁና:: **እርሱ** በዚህ የለም::
56. **እግዚአብሔር** ቸር መሆኑን ቀምሳችሁ እዩ:: **እርሱን** መጠጊያ የሚያደርግ ምንኛ ቡሩክ ነው::

57. **እግዚአብሔር** ጥበብን ይሰጣልና:: **እርሱ** ለቅኖች ድልን ያከማቻል::

58. **የእግዚአብሔር** ቃል እንከን የለበትም:: **እርሱ** ለሚታመኑበት ጋሻቸው ነው::

59. **እግዚአብሔር** የሚንገዳገዱትን ሁሉ ይደግፋል:: የሁሉ ዐይን **እርሱን** በተስፋ ይጠብቃል::

60. **እግዚአብሔር** መሓሪና ይቅር ባይ ነው:: **እርሱ** ሁልጊዜ በደልን አይከታተልም ::

Appendix B: sample independent & hidden personal pronouns referring nouns which are definite

1. **መስተዳድሩ** ከፍተኛ ውጤት ላስመዘገቡ ተማሪዎች ሽልማት **ሰጠ**
2. **አስተማሪው** ለተማሪዎቹ መጽሀፍ **ሰጣቸው**
3. **አትክልቱ** በደንብ **ለማ**
4. **ቢሮው** ፈቃድ የተሰረዘባቸውን ነጋዴዎች ዝርዝር **አወጣ**
5. **ሰራተኛዋ** ዝናብ ያረጠበውን ልብስ ቤት ውስጥ **አሰጣች**
6. **ዘበኛው** አሰሪው ስለፈቀደለት የማታ ትምህርት **ተመዘገበ**
7. **ፕሬዚዳንቱ** አደጋው ህዝብ እንዳሳዘነ **ገለጹ**
8. **ፋብሪካው** 85ነጥብ7ሚሊየን ብር የተጣራ ትርፍ **አገኘ**
9. **ልጅቷ** ትልቅ ወንድሟ እንዲያስጠናት ከፍል ውስጥ **ነገረችው**
10. **ሽምግሌው** ጥርሳቸው ስላለቀ ብዙ ጥሬ መብላት **አቆሙ**
11. **ድርጅቱ** በ33ነጥብ5ሚሊየን ብር የመንገድ ግንባታዎችን **እያካሄደ ነው**
12. ተዘዋዋሪ **ችሎቱ** 550 እስረኞች እንዲፈቱ **በየነ**
13. **እንስሶቹ** እንደ አሳ **ይዋኛሉ**
14. **ነጋዴው** ትንሽ የደብረብርሀን በግ **ገዛ**
15. **አዛዡ** ወታደሮቹ ወደ ግቢያቸው እንደገቡ በምስጋን **ተቀበሏቸው**
16. **ወፎቹ** በቀለ ያፈሰሰውን ስንዴ በችኮላ **ለቀሙ**
17. **አሮጊቷ** ሌቦች አምና የገደሉትን ብቸኛ ልጃቸውን **አስታወሱ**
18. **ገቢያተኞቹ** እኛ የቆምንበትን አውላላ ሜዳ ወድያው **ሞሉት**
19. **ሊቀመንበሩ** የህዝብ ገንዘብ **ሰረቀ**
20. **አባትየው** ልጁ አንደኛ ስለወጣ የወርቅ ሰዓት **ሸለመው**
21. **ልማትፈንዱ** 16ሚሊየን ነዋሪዎችን ተጠቃሚ ማድረጉን **ገለጸ**
22. **ጋዜጠኛው** ንጉሱ በአሜሪካ የተደረገላቸውን አቀባበል በደንብ **ዘገቡ**
23. **ዲስትሪክቱ** በ36ሚሊየን ብር የመንገድ ጥገና **አከናውኗል**
24. **ቦርሳው** ብዙ እቃ በውስጡ ስለያዘ በጣም **ከበደን**
25. የጤና ተቋማቱ ግንባታ ከ10 ወራት በኋላ ሲጠናቀቅ 65ሽህ ለሚጠጉ ነዋሪዎች አገልግሎት ለመስጠት እንደሚያስችሉም **ተወካዩ አስታውቀዋል** ።
26. የአለም አቀፉ የፓርላማ ህብረት የኢትዮጵያ ፓርላማ የአዳድስ ቴክኖሎጂዎች ተጠቃሚ እንድሆን ለማስቻል ከአፈ ጉባኤ አምባሳደር ተሾመ ጋር ውይይት መጀመራቸውን **ፕሬዚዳንቱ አስረድተዋል** ።

27. የዘርፉን ስፖርት ለማስፋፋትና ለማሳደግ እንድሁም የበለጠ ውጤታማ ለመሆን ሁሉም አካላት እጅ ለእጅ ተያይዘው መስራት እንደሚገባቸውም **ፕሬዚዳንቱ አስገንዝበዋል ።**
28. **ድርጅቱ** በመላው አገሪቱ የስርጭት አድማሱን ለማስፋፋት **አቅዷል ።**
29. **ድርጅቱ** ዛሬ ለኢትዮጵያ ዜና አገልግሎት እንዳስታወቀው በየእለቱ ለ12 ሰአታት የነበረውን የቴሌቪዥን ስርጭት ወደ 15 ሰአት ከፍ እንድል **አድርጓል ።**
30. **ድርጅቱ** የአጭር ጊዜ የቴሌቪዥን የአየር ሰአት ድልድል **አደረገ**
31. ***መልአኩም*** ሴቶቹን እንዲህ አላቸው አትፍሩ የተሰቀለውን ኢየሱስን እንደምትፈልጉ አውቃለሁና:: **እርሱ** በዚህ የለም::
32. ፕላጦስ ሁኔታው ሽብር ከማስነሳት በስተቀር ምንም ፋይዳ የሌለው መሆኑን ተመልክቶ እኔ ከዚህ ሰው ደም ንጹሕ ነኝ በማለት ውሃ አስመጥቶ በሕዝቡ ፊት እጆቹን ታጠበ:: **ሕዝቡም** በሙሉ ደሙ **በእኛና** በልጆቻችን ላይ ይሁን ብለው መለሱ
33. **ወንድሞቹም** አባታቸው **ከእነርሱ** አብልጦ የሚወደው መሆኑን ሲያዩ ዮሴፍን ጠሉት
34. **ሴቲቱም** የዛፉ ፍሬ ለመብል መልካም እንደሆነ ባየች ጊዜ ከፍሬው ወስዳ በላች:: **ከእርሷም** ጋር ለነበረው ለባሏ ሰጠችው እርሱም በላ

Appendix C: sample independent & hidden personal pronouns referring nouns which are recent

1. የጋምቤላ ብሄራዊ ፖርክን ለማሻሻል 110ሺ **ብር ተመደበ**
2. እነዚያ **ጌሾዎች ተቆረጡ**
3. ከአስቴር የበለጠች ቆንጆ **ልጅ** ከጎጃም **መጣች**
4. ሁለት ጣሳ የጉበስ **ጠላ** ከገቢያ **ተገዛ**
5. አልቃይዳ በሶማሊያ እንደገና እንዳይደራጅ **አሜሪካ** ቅኝቷን **አጠናክራለች**
6. በደንቆሮ ደን የቀይ ቀበሮ **ቁጥር እየጨመረ** ነው
7. ሁለት ሊትር ንጹህ የማር **ጠጅ ተሸጠ**
8. አሮጌ ሚሳይል በሞቃዲሾ ፈንድቶ 14 **ሰዎች ሞቱ**
9. የጎንደር ከተማ ዘላቂ የመጠጥ ውሃ **ፕሮጀክት** ዛሬ **ተመረቀ**
10. ሚሳይል በሞቃዲሾ ፈንድቶ 14 **ሰዎች ሞቱ**
11. ካሳ በግ ያረደበነት **ቢላዋ** በጣም **ስለዛገ ተጣለ**
12. የሀገሪቱ ባለስልጣናት የህዘን **መግለጫ መልእክት ቀጥሏል**
13. ካሳ የገዛው **በግ** ትናንት **ጠፋ**
14. የሀገሪቱ ባለስልጣናት መግለጫ **መልእክት ቀጥሏል**
15. የኢሃዴግ 4ኛ **ጉባኤ ተከፈተ**
16. በአውሮፓ ህብረት የኤርትራ **አምባሳደር** ከሃላፊነት **ሊለቁ** ነው
17. ካሳ የተከላቸው ትልልቅ ሁለት እግር **ጌሾዎች ተቆረጡ**
18. በቡግና ወረዳ በርካታ የቱሪስት **መስህቦች አሉ**
19. ክፍል ውስጥ የሚገኙት ጥቂት **ተማሪዎች** አሁን **ወጡ**
20. የካሳ የወተት **ላሞች** በጣም **ታመሙ**
21. በቦረና የኦህዴድ ተሃድሶ **ውይይት ተጀመረ**
22. ካሳ የገዛው የትናንቱ **በግ ታረደ**
23. ካሳ የገዛው የሱፍ **ኮት** በጣም **ያምራል**
24. ያዕቆብም ምግቡን **ለአባቱ** አቀረበለት:: **እርሱም** በላ::
25. የሰው መንገድ **በእግዚአብሔር** ፊት ግልጥ ነውና:: **እርሱ** መሄጃውን ሁሉ ይመረምራል::
26. በፍጹም ልብህ **በእግዚአብሔር** ታመን:: በመንገድህ ሁሉ **እርሱን** ዕወቅ::
27. ምድር ሆይ **በጌታ** ፊት ተንቀጥቀጪ :: **እርሱ** ዐለቱን ወደ ኩሬ ለወጠ::

28. ኢትዮጵያ የዘንድሮውን የምስራቅ አፍሪካ አገር አቋራጭ ሻምፒዮና የፊታችን እሁድ እንደምታስተናግድ **ፌደሬሽኑ ገለፀ ።**
29. በተከሳሹ የባህላዊ ህክምና ፍቃድ ወረቀት ላይ የራሱ ፎቶ ግራፍ ቢኖርም መታወቂያው የሌላ ሰው ስም ሆኖ በመገኘቱ ቅጣቱ ሊከብድ መቻሉን **ዳኛዋ አስረድተዋል ።**
30. ሚስተር አልመርትን ይተኳቸዋል ተብለው የሚጠበቁት የውጭ ጉዳይ ሚኒስትሯ ትስፒ ሌቭኒ መሆናቸውን **ቢቢሲ ዘግቧል ።**

Appendix D: sample hidden personal pronouns referring personal pronouns

1. **እኔ** ካሳ የገዛውን ትንሽ በግ ትላንት **አየሁት**
2. **እኔ** ልብስ **ለበኩ**
3. **እኔ** የክፍል ጓደኛዬ የተኛበትን አልጋ ከጎረቤት **ተዋስኩ**
4. **እኛ** ስራ **ሰራን**
5. **እኔ** ልብሴ ዛሬ ስለታጠበ አሮጌ ልብስ **ለበኩ**
6. **እኔ** አልጋ **ተዋስኩ**
7. **እኔ** በግ **አየሁ**
8. **እሷ** ጎበዝ ተማሪ **ትሆናለች**
9. **እሷ** እሱን **ሰደበችው**
10. **እሱ** አባቱን **ይመስላል**
11. **እኛ** አስተማሪው የሰጠንን የቤት ስራ ክፍል ውስጥ **ሰራን**
12. **እነርሱም** እርሱ ወደ **እነርሱ** ሲመጣ በሩቅ አዩት::

Appendix E: sample independent & hidden personal pronouns referring nouns mentioned most

1. **ዮሴፍ ወንድሞቹን ፍለጋ ሄደ። ወንድሞቹም ዮሴፍ ወደ እነርሱ ሲመጣ በሩቅ አዩት ።**
2. **ኢየሱስ በዮሐንስ ለመጠመቅ ከገሊላ ወደ ዮርዳኖስ መጣ። ዮሐንስ ግን እኔ ባንተ መጠመቅ ሲያስፈልገኝ እንዴት አንተ ወደ እኔ ትመጣለህ ብሎ ተከላከለ ።**

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: Temesgen Dawit

Signature: _____

Date: _____

Confirmed by advisor:

Name: Yaregal Assabie (PhD)

Signature: _____

Date: _____

Place and date of submission: Addis Ababa, April, 2014.

Abstract

Anaphora resolution is the process of finding an entity which points backward to a word or phrase that has been introduced with more descriptive phrase in the text than the entity or expression which is pointing back. An entity referring back is called *anaphor*, whereas the word or phrase being referred is called *antecedent*. Anaphora resolution is used as a component in NLP applications like machine translation, information extraction, question answering and others to increase their effectiveness. Building complete anaphora resolution systems that incorporate all linguistic information is complex and still not achieved because of the different nature of languages and their complexities. In the case of Amharic language, it is even more complex because of its rich morphology. In addition to independent anaphors, unlike other languages like English, Amharic language has anaphors embedded inside words (hidden anaphors).

In this work, we have proposed Amharic anaphora resolution model using *knowledge poor* anaphora resolution approach. The approach uses low levels of linguistic knowledge like morphology to build anaphora resolution systems avoiding the need of complex knowledge like semantic analysis, world knowledge and others. The proposed model takes Amharic texts as input and preprocesses to tag the texts with word classes and various chunks. Anaphors, both independent and hidden, and antecedents are identified from the preprocessed dataset. The model deals with both intrasentential and intersentential type of anaphors. Finally, the resolution process uses constraint and preference rules to identify the correct antecedent referred by the anaphor.

To evaluate the performance of the model, Amharic texts are collected from Walta Information Center (WIC) and Amharic Holy Bible and used as datasets. The collected dataset was divided into training and testing datasets based on 10-fold cross validation technique. Based on the collected dataset, we achieved a success rate of 81.79% for resolution of hidden anaphors whereas an accuracy of 70.91% was obtained for resolution of independent anaphors.

Keywords: Amharic anaphora resolution, knowledge poor anaphora resolution approach, hidden anaphors.