ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

DESIGNING AN INFORMATION EXTRACTION SYSTEM FOR AMHARIC VACANCY ANNOUNCEMENT TEXT

SINTAYEHU HIRPASSA

JUNE 2011

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

DESIGNING AN INFORMATION EXTRACTION SYSTEM FOR AMHARIC VACANCY ANNOUNCEMENT TEXT

SINTAYEHU HIRPASSA

JUNE, 2013

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

DESIGNING AN INFORMATION EXTRACTION SYSTEM FOR AMHARIC VACANCY ANNOUNCEMENT TEXT

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

SINTAYEHU HIRPASSA

JUNE 2013

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION SCIENCE

DESIGNING AN INFORMATION EXTRACTION SYSTEM FOR AMHARIC VACANCY ANNOUNCEMENT TEXT

By

SINTAYEHU HIRPASSA

Name and signature of Members of the Examining Board

Name	Title	Signature	Date
_____	Chairperson,	_____	_____
_____	Advisor(s),	_____	_____
_____	Advisor(s),	_____	_____
_____	Examiner	_____	_____

ACKNOWLEDGMENT

All the praise goes to almighty God, who gives me the strength from beginning to the end of this research work. Next my best gratitude goes to my advisor Ato Ermias Abebe for his continuous support in materials and invaluable comments. He spent his precious time in commenting my work and showing me the right directions which I found very important for my thesis.

I owe a considerable debt to Ato Yibekal Getahun database administrator in Reporter Newspaper for providing me the necessary data that is important for training and testing of the prototype developed. I extend my thanks to all Instructors from Institute of Language Studies, Department of Amharic Language for their support in studying Amharic language.

My special thanks goes to my brothers, Fekadu for making extremely supportive and encouraging in difficult times to become my life prettier, and also, for passing all those hardships I had to go through easier.

Finally, I would like to express my sincere thanks to all my family and very close friends who has been working day in day out with me whenever in need.

TABLE CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS & ABBREVIATIONS

AVAT: Amharic vacancy announcement texts
AVATIES: Amharic Text Information Extraction
GATE: General Architecture for Text Engineering
HMM: hidden Markov Model
IE: Information Extraction
IR: Information Retrieval
NER: Named Entity Recognition
NLP: Natural Language Processing
POS: Part-of-Speech

ABSTRACT

The number of Amharic documents on the Web is increasing as many newspaper publishers started providing their services electronically. The unavailability of tools for extracting and exploiting the valuable information from Amharic text, which is effective enough to satisfy the users has been a major problem and manually extracting information from a large amount of unstructured text is a very tiresome and time consuming job, this was the main reason which motivate the researcher to engage in this research work.

The overall objective of the research was to develop information extraction system for the Amharic vacancy announcement text. The system was developed by using Python and visual basic programming language and rule-based technique was applied to address the problem of automatically deciding the correct candidate texts based on its surrounding context words. 116 Amharic vacancy announcement texts which contain 10,766 words were collected from the "Ethiopian reporter" newspaper published in Amharic twice in week.

For this study, nine candidate texts are selected from Amharic vacancy announcement text, these are organization, position, qualification, experience, salary, number of people required, work agreement, deadline and phone number. The experiments have been carried out on each component of a system separately to evaluate its performance on each components, this helps us to identify drawbacks and give some clue for future works.

The experimental result shows, an overall F - measure of 71.7% achieved. In order to make the system to be applicable in this domain which is Amharic vacancy announcement, further study is required like incorporating additional rules, improving the speed of the system by modifying the algorithm, a well designed user interface and integrating other NLP facilities.

CHAPTER ONE
INTRODUCTION

1.1. GENERAL BACKGROUND

Rapid developments in Information and Communication Technology are making available huge amount of data and information. Much of these data is in electronics forms (like more than billion documents in the World Wide Web). Usually these data are unstructured or semi-structured and can generally be considered as a text database. Likewise, the recent decades witnessed a rapid proliferation of Amharic textual information available in digital form in a myriad of repositories on the Internet and intranets. As a result of this growth, a huge amount of valuable information, which can be used in education, business, health and other many areas are hidden under unstructured representation of the textual data and is thus hard to search in. This resulted in a growing need for effective and efficient techniques for analyzing free-text data and discovering valuable and relevant knowledge from it in the form of structured information, and led to the emergence of Information Extraction technologies.

Information Extraction (IE) is one of the NLP applications that aim to automatically extract structured factual from unstructured text. Riloff [2] discusses, the task of automatic extraction of information from texts involves identify a predefined set of concepts and deciding whether a text is relevant for a certain domain, and if so extracting a set of facts from that text.

IE has three different components regardless of the language and domain on which it is developed for. The components are linguistic preprocessing, learning and application, and post processing. Linguistic preprocessing uses different tools to make the natural language texts ready for extraction. The learning and the application component learns a model and extract the required information from the preprocessed text.

In the last component the semantic post processing assign the extracted information into their predefined attribute category and manages the normalization and duplication problem with the extracted data [5].

In principle, designing IE has two approaches: (1) the learning approach, and (2) the Knowledge Engineering approach. For systems or modules using learning techniques an annotated corpus of domain relevant texts is necessary. This approach calls for someone who has enough knowledge about the domain and the tasks of the system to annotate the texts appropriately. The annotated texts are the input of the system or module, which runs a training algorithm on them. Thus, the system obtains knowledge from the annotated texts and can use it to gain desired information from new texts of the same domain.

The Knowledge Engineering (KE) approach needs a system developer, who is familiar with both the requirements of the application domain and the function fof the designed IE system. The developer is concerned with the definition of rules used to extract the relevant information. Therefore, a corpus of domain-relevant texts will be available for this task [9].

IE is quite different from IR. An IR system finds relevant texts that is based on a query and presents them to the user. An IE application analyzes texts and presents only the specific information from it that the user is interested in.
IE systems are more difficult and knowledge-intensive to build, and are to varying degrees tied to particular domains and scenarios. It is also more computationally intensive than IR. In applications where there are large text volumes IE is potentially much more efficient than IR because of the possibility of dramatically reducing the amount of time people spend reading texts [4].

During the last ten years, IE has become an increasingly researched field. As [2] stated, “unfortunately, during this time most of the known IE systems have been invented for texts written in the English language. In comparison to the success registered for English IE systems for most of other languages are still lacking essential components”.

Depending on the number of native speakers, prosperity of countries, and the need for natural language processing capabilities, as well as due to the complexity of certain languages, IE systems are uniquely designed for individual languages.

1.2. STATEMENT OF THE PROBLEM

With the popular use of the World Wide Web as global information system a number of newspapers are already flourishing online. Likewise, in Ethiopia most of Amharic newspaper publishers are providing their publications online. Among the well known newspapers in Ethiopia, the "Ethiopia Reporter" is the one. It appears twice a week with contents such as news, politics, science and technology, sport, business, vacancy and social. The newspaper presents different vacancies of organizations in structured, unstructured and semi-structured forms. Cowie and Wilks [3] noted, manually extracting information from such an often unstructured or semi-structured text is a very tiresome and time consuming job. Thus, getting the right information for decision making from existing abundant unstructured text is a big challenge.

In addition, the unavailability of tools for extracting and exploiting the valuable information which is effective enough to satisfy the users for Amharic language has also been a major problem. It is hoped that the availability of an IE tool can ease this information searching process.

IE unlike the other research domains is language and domain dependent [38]. The IE system developed for English and in this specific domain may not work for Amharic language even if its domain is similar. There are different, language specific, issues which may not be handled by the system developed for English. This is due to the reason that IE system has to be trained about the different nature of the language and the domain for which they are developed for.

To the best knowledge of the researcher, the work of Tsedalu [21] has only been one research conducted on Amharic IE system. The work is also limited in extracting numeric and nominal data from the Amharic news text. News texts that are about a single issue are only considered and extraction of relationship between entities is out of the scope of this research work. In addition, even if the language is similar with this research its domain is different. This system may not fully handle the concerns that are viewed in information extraction from AVAT. Thus, this and other reasons initiate the researcher to engage in research to design IE for AVAT.

This study has attempted to answer the following research questions:

1. What approaches should be followed in designing an Amharic IE system that identifies useful information from vacancy announcement?
2. What algorithms are suitable for automatic Amharic IE?
3. What model ought to suppose to design Amharic IE?

1.3. OBJECTIVE OF THE STUDY

1.3.1 GENERAL OBJECTIVE

The general objective of this study is to design information extraction system for AVAT.

1.3.2. SPECIFIC OBJECTIVES

The specific objectives of the research are:

- To review word categorization and character representation in Amharic language.
- To build up an architecture for IE for AVAT.
- To develop suitable approaches and algorithms for IE
- To develop a prototype system that demonstrates the potentials of the Amharic IE system.

- To evaluate the performance and usability of the prototype developed for Amharic information extraction

1.4. METHODOLOGY

1.4.1. STUDY DESIGN

The design of this research is experimental. In this study different activities were involved. Identifying the problem in the area of AVAT was the starting point of the study. To address the problem IE system is designed and implemented. It is obvious that, testing is mandatory for any type of system once it designed, to check its applicability and to evaluate its performance. In the same way, the system which is designed in this study is tested and evaluated based on the test dataset.

1.4.2. LITERATURE REVIEW

In order to have a better understanding of in IE and design a system for Amharic language, different local and global researches were thoroughly reviewed. Literature such as journals, articles, proceeding, papers and books were reviewed for achieving the objective of this research.

1.4.3. DATA SOURCES AND DATA PREPARATION FOR THE EXPERIMENT

The researcher collected different AVAT that were required for training and testing the system from the "Ethiopian Reporter" newspaper published in Amharic twice in week. For the purpose of this study, 116 AVATs that contain in general 10,766 words were selected purposefully with different range of vacancy announcements. There dissimilarity is based on the organization of who is posted the vacancies and the type of vacancies. The newspaper was chosen as a data source since it has large collection of AVAT in its database.

After the raw AVAT were collected, different data preprocessing tasks were undertaken (such as tokenization, normalization, transliteration) and, gazetteer was prepared.

UNDERSTANDING OF DOMAIN LANGUAGE

The different facts about Amharic language like the word categorization, character and number representation and other language specific issues that are important for the research work have been analyzed and presented. It helps to understand the nature of the language with regard to information extraction.

1.4.4. DESIGN AND IMPLEMENTATION OF AVATIES

The designing phase contains the document preprocessing, learning and extraction, and post processing as the three main components. In order to develop a prototype system, different appropriate tools have been selected and used.

The different data preprocess IE components, such as Tokenizer, Normalizer, Transliterator, etc, which are mostly language specific algorithms are developed using python programming language. This programming language was employed for developing candidate text selector and tagger and candidate text extractor. The main reason that the python programming language is used is for the familiarity of the language with the researcher.

The POS which is developed by Gebrekidan [36] is used as one of the features in IE component. Also, Microsoft SQL server 2008 was used to store extracted candidates and Visual Basic programming language was used for the development of user interface, which helps a user to interact with the system and access data from database.

1.5. APPLICATION OF RESULTS AND BENEFICIARIES

Nowadays most people use online newspapers as a source of information for vacancy announcements. Thus, those who use newspapers and websites for job search are the main beneficiaries of this study. It will help them save their time in searching detailed information about the jobs that are posted in unstructured format. It also helps them access facts or details easily.

The system will also have great significance for publishers of newspapers as it will help to provide vacancy news in attractive and structured fashion. It can also keep them from committing errors while in changing unstructured AVAT in to structure and also will have a tremendous effect in enhancement of their day to day activities.

Beside to this, the extracted structured information from unstructured text can be used as an input for other applications such as question and answering application system and etc.

At the end, it is hoped that, this study will serve as tipping point for other researchers to focus much on this research issue.

1.6. SCOPE AND LIMITATIONS OF THE STUDY

The task of designing information extraction system requires a very intensive knowledge in natural language processing. The main limitation while processing the study is the unavailability of enough corpus and word categories for natural language processing for the domain. This would set a constraint on amount of rule generation.

A full-fledged information extraction system will require a number of NLP tools such as Sentence Parser, Part of Speech tagger (POS), Named Entity Recognizer (NER), Co-reference Resolution and others. Even though some of the NLP systems for Amharic language have been done by other researchers, they are not publicly available.

Having these limitations in mind, the researcher tried to design a rule-based IE system only for a specific domain, which is AVAT, and this study was confined only to extract organization, job title, required qualification, work experience, salary, number of people required, job agreement and deadline data from AVAT. Information extraction of other information type from the AVAT is out of the scope of the study.

1.7. ORGANIZATION OF THE STUDY

The thesis is organized into six chapters. The first chapter of the thesis contains background, statement of the problem, objectives, and methodology of the research. Chapter 2 discusses the different issues in IE and the related subject areas as literature review. Also this Chapter lays the foundation in understanding what an IE system comprises of, what approaches are used, and the different components which are required by the IE system. The last part of the chapter, discusses related works on IE systems in different languages and on different domains.

In chapter 3, a discussion is made about Amharic language with regard to IE. Many language specific issues such as the writing system and language structure are presented. Chapter four is devoted to discussing the architectural and design issues of the system, the main components of our system, their functional operation and the specific sub-component of each component are briefly discussed. In this chapter it also discusses the main implementation issues of our IE system, the algorithms and techniques used to develop the system successfully. Result and performance evaluation of the system is presented in chapter five. Finally, conclusion and recommendations for further study is forwarded.

CHAPTER TWO
LITERATURE REVIEW

2.1. INTRODUCTION

Now a day, an increasing amount of information is available in the form of electronic documents. This makes it nearly impossible to manually search, filter and choose which information one should use for his/her own purpose [20].

Different scholars tried to develop different information management systems so that the drawing of summarized and relevant information from an ocean of information can be facilitated and the right information for decision making can be acquired. Among the different solution to the problems are Information Retrieval (IR), Information Extraction (IE), Question Answering, Text Summarization and Text Categorization [21].

In this section we will describe the requirements and components of IE systems as well as present various approaches for building such systems. Then, we will present important methodologies and systems for IE systems.

The related NLP fields are also reviewed and presented in order to see their similarity and difference with IE. Evaluation standards for the performance of IE system which are used for the evaluation purpose are also presented in this chapter.

2.2. INFORMATION EXTRACTION (IE)

IE has become an important notion to address the problem of information overload by locating the target phrases from document and transforms them in to structured representation.

As it is defined by Eikvil [15] “it is the task of locating specific pieces of data from a natural language document, a particularly useful sub-area of natural language processing (NLP). In IE, the data to be extracted from a natural language text is given by a template may be either one of a set of specified values or strings taken directly from the document”

Mooney and Califf [16] also defines IE as “IE is a form of shallow text processing that locates a specified set of relevant items in a natural language document, transforming unstructured text into a structured database”. Systems for this task require significant domain-specific knowledge. So generally, IE is the process of extracting relevant and factual data from unstructured or free text.

IE usually uses NLP tools, lexical resources and semantic constraints for better efficiency [21]. The General Architecture for Text Engineering (GATE) which is the widely known open source software system for computations related to natural language defines IE as a system which analyses unstructured text in order to extract information about pre-specified types of events, entities or relationships.

According to Wilks and Brewster [8], the requirement of templates and bundling domain and corpus specific information with the IE techniques are two major challenges on IE.

2.3. BUILDING INFORMATION EXTRACTION SYSTEMS

At this point, we shall turn our attention to what is actually involved in building IE systems. Before discussing in detail the basic parts of an IE system, we point out that there are two basic approaches to the design of IE systems, which we label as the Knowledge Engineering Approach and the Automatic Training Approach.

I. KNOWLEDGE ENGINEERING APPROACH

The Knowledge Engineering Approach is characterized by the development of the grammars used by a component of the IE system by a "knowledge engineer," i.e. a person who is familiar with the IE system, and the formalism for expressing rules for that system, who then, either on his own, or in consultation with an expert in the domain of application, writes rules for the IE system component that mark or extract the sought after information.

Typically the knowledge engineer will have access to a moderate-size corpus of domain-relevant texts (a moderate-size corpus is all that a person could reasonably be expected to personally examine), and his or her own intuitions [1]. It is obviously the case that the skill of the knowledge engineer plays a large factor in the level of performance that will be achieved by the overall system. In addition to requiring skill and detailed knowledge of a particular IE system, the knowledge engineering approach usually requires a lot of labor as well [1].

Building a high performance system is usually an iterative process whereby a set of rules is written, the system is run over a training corpus of texts, and the output is examined to see where the rules under and over generate. The knowledge engineer then makes appropriate modifications to the rules, and iterates the process [1]. Thus, the performance of the IE system depends on the skill of the knowledge engineer.

II. AUTOMATIC TRAINING APPROACH

The Automatic Training Approach is quite different. Following this approach, it is not necessary to have someone on hand with detailed knowledge of how the IE system works, or how to write rules for it.

It is necessary only to have someone who knows enough about the domain and the task to take a corpus of texts, and annotate the texts appropriately for the information being extracted.

Typically, the annotations would focus on one particular aspect of the system's processing. For example, a name recognizer would be trained by annotating a corpus of texts with the domain-relevant proper names.

A co-reference component would be trained with a corpus indicating the co-reference equivalence classes for each text. Once a suitable training corpus has been annotated, a training algorithm is run, and resulting in information that a system can employ in analyzing novel texts. Another approach to obtaining training data is to interact with the user during the processing of a text. The user is allowed to indicate whether the system's hypotheses about the text [4, 9]. The above mentioned approaches for IE can be applied on the free text or semi structured or structured text which is used as an input for IE system [16].

***Free text*:** is unstructured collection of text. It can't be easily managed as it doesn't have the structure or any predefined format in order to manage it by using computers. The natural language components are applied in order to manage extraction from the free text [21].

Semi Structured Text: is a data which is not in the form of tuples like structured text and is different from free texts which rather exist in between the two. The information in the form of HTML tags is semi structured text [21].

Structured Text: is textual information which exists in a database or file following a predefined and strict format. Such information can easily be extracted by using the format description as it has a known format [21].

IE approaches supported on supervised machine learning technique are divided in to the following three categories [21]

I Rule learning

II Linear separators

III Statistical learning

I Rule Learning

This approach is based on a symbolic inductive learning process. The extraction patterns represent the training examples in terms of attributes and relations between textual elements.

Some IE systems use propositional learning (i.e. zero order logic), for instance, Auto Slog-TS and CRYSTAL, while others perform a relational learning (i.e. first order logic), for instance WHISK and SRV. This approach has been used to learn from structured, semi-structured and free-text documents [21].

II Linear Separators

In this approach the classifiers are learned as sparse networks of linear functions (i.e. linear separators of positive and negative examples). It has been commonly used to extract information from semi-structured documents. It has been applied in problems such as extraction of data from job ads, and detection of an e-mail address change [21].

In general, the IE systems based on this approach present an architecture supported on the hypothesis that looking at the words combinations around the interesting information is enough to learn the required extraction patterns. [21].

III Statistical Learning

This approach is focused on learning Hidden Markov Models (HMMs) as useful knowledge to extract relevant fragments from documents [21].

These IE systems also differ from each other in the features that they use. Some use only basic features such as token string, capitalization, and token type (word, number, etc.).

In addition, others use linguistic features such as part-of-speech, semantic information from gazetteer lists, and the outputs of other IE systems (most frequently general purpose named entity recognizers).

A few systems also exploit genre-specific information such as document structure. In general, the more features the system used, the better performance it could achieve. One of the most successful machine learning methods for IE is Support Vector Machine (SVM), which is a general supervised machine learning algorithm. It has achieved state-of-the-art performance on many classification tasks, including named entity recognition.

2.4. ARCHITECTURE OF INFORMATION EXTRACTION SYSTEM

Different scholars use different steps for designing extracting information system for different language and different domain. The research work in [1] mainly categorizes IE in to six different tasks.

- Part-of-Speech (POS) Tagging
- Named Entity Recognition (NER)
- Syntax Analysis
- Co-references and Discourse Analysis
- Extraction Patterns
- Bootstrapping

I. Part-of-speech tagging (POS)

It is the act of assigning each word in sentences of tag that describes how that word is used in the sentences. That means POS tagging assigns whether a given word is used as a noun, adjective, verb, etc.

As Pal and Molina [23] acknowledges, one of the most well-known disambiguation problem is POS tagging, because many words are ambiguous: they may be assigned more than one POS tag (for example, the English word round may be a noun, an adjective, a preposition or an adverb, or a verb).

POS tagger finds the possible tags or lexical category for each word provided that the word is in a lexicon and guess possible tags for unknown words. It also chooses possible tag for each word that is ambiguous in its part-of-speech. If certain word is assigned more than one tag, this means that the word can have different meanings or function in different context.

According to Antonio [29], there are two approaches to automatic POS tagging: rule-based approaches use linguistic knowledge to formulate simple rules that assign a part of speech to an ambiguous word using context information; statistical approaches of which hidden Markov models trained using the expectation-maximization algorithm are the standard model) use the statistics collected from ambiguously or unambiguously tagged texts to estimate the likelihood of each possible interpretation of a sentence or text portion so that the most likely disambiguation is chosen.

II. Named Entity Recognition (NER)

Named entities are one of the most often extracted types of tokens during extracting information from documents. Named entity recognition is classification of every word in a document as being a person-name, organization, location, date, time, monetary value, percentage, or “none of the above”. Some approaches use a simple lookup in predefined lists of geographic locations, company names, person names and name of animals and other things from the gazetteers, while some others utilize trainable Hidden Markov Models to identify named entities and their type.

For example the NER takes the following AVAT recognize the named entities and numbers which will be used as attributes for the predefined database slot

አ/ማኅበራችንከዚህበታችበተመለከተውየሥራመደብአመልካቾችንአወዳድሮ በኮንትራት ለመቅጠር ስለሚፈልግ የትምህርትና የሥራ ልምድ ማስረጃዎቻችሁን በመያዝ በ10 ተከታታይ የሥራ ቀናት ውስጥ ሰው ሃብት ሥራ አመራር ቡድን ቢሮ ቁጥር 104 እየቀረባችሁ መመዝገብ የምትችሉ መሆኑን እናስታውቃለን። የሥራመደቡ የሕግባለሙያ ተፈላጊችሎታ ከታወቀ ዩኒቨርስቲ በሕግ የመጀመሪያ ዲግሪ በመደበኛ የትምህርት ክፍለ ጊዜ የጨረሰና ከምረቃ በኋላ 6 ዓመት የሥራ ልምድ የቅጥር ሁኔታ ኮንትራት ደመወዝ በስምምነት አድራሻ፡- ንፋስ ስልክ የቀድሞው ኢሜክስኮ ቅጥር ግቢ 1ኛ ፎቅ ቢሮ ቁጥር 104 ስ.ቁ 011-42-38-64 ወይም 0114-42-47-77 የውስጥ መስመር 525/253 ኮሜት ትራንስፖርት አክሲዮን ማኅበር

The words which are names and numbers that represent different thing will be extracted like ንፋስ ስልክ, ኮሜት ትራንስፖርት አክሲዮን ማኅበር , 104, ሕግ ባለሙያ, ሕግ, 6, ዲግሪ, ሰው ሃብት ሥራአመራር ,ዩኒቨርስቲ which are the named entity attributed in the text which represent different things.

III. Syntax analysis

In contrast to POS tagging, syntax analysis, also called syntax parsing, looks beyond the scope of single words. During syntax analysis we attempt to identify syntactical parts of a sentence (verb group, noun group and prepositional phrases) and their functions (subject, direct and indirect object, modifiers and determiners). Simple sentences, consisting, for instance, of a main clause only, can be parsed using a finite state grammar. Simple finite state grammars are often not sufficient to parse more complex sentences, consisting of one or more subordinate clauses in addition to the main clause, or containing syntax structures, such as prepositional phrases, adverbial phrases, conjunction, personal and relative pronouns and genitives in noun phrases.

Using finite state grammars in such cases may result in errors. Instead, those cases are handled by statistically founded methods which have to be trained with training text corpora.

The important decision to be made for syntax analysis is basically the same as for named entity recognition and POS tagging.

We have to decide what kind of parsing is to be employed: more robust shallow techniques, or deep complex syntax analysis [24]

IV. Co-references and Discourse Analysis

It is a process of finding multiple references to the same object in a text. It refers to the task of identifying noun phrases that refer to the same extra linguistic entity in a text. This is especially important since the same thing about a single entity is expressed in different sentences using pronouns [1].

V. Extraction Patterns

The resulting output of IE consists of single data items filled into the slots of data tuple templates. The data tuples populate the result database, one tuple for each relevant document of the input text corpus. The data items are pieces of information which have to be located in the text. Extraction patterns are used for this task.

An extraction pattern is a text pattern which matches a certain token and its surrounding context. [1].

VI. Bootstrapping

As Johannes [1] notes that, newer systems use various bootstrapping algorithms to improve the results of the pattern matching, or do unsupervised named entity recognition. Some systems require a test corpus to evaluate the results of the pattern matching and bootstrapping process.
During the bootstrapping the following steps are iterated:

1. Apply all seed patterns on the whole text corpus.
2. Split the text corpus into two categories, so that one category contains all relevant texts in which one or more seed patterns scored and the other category contains all the other texts.

3. Score all the patterns gained from the text corpus based on their density of distribution in relevant documents in comparison to their density of distribution in all texts.
4. Use the highest scoring patterns to generate concept classes by merging those pairs which appear in the correlated text.

In other hand, as Tsedalu [17] citied the work of Yannick and Versley, IE tasks are categorized in to five different and independent components.
As IE activity can be a very complex task decomposing it into different task is advantageous. The main advantage of decomposition is, it helps to choose the techniques and algorithms that suit each task, and to debug an IE program easily.

The considered tasks in the survey are:

- Segmentation,
- Classification,
- Association,
- Normalization and
- Co-Reference Resolution.

I. Segmentation

The Segmentation task divides the text into atomic elements, called segments or tokens. Even though this task is simplified for Western languages due to the existence of white spaces separating words, there are some cases in which simple white space separation may not be enough. Usually, segmentation for these cases is performed using rules that show how to handle each case. The major problems related to this task can be found in oriental languages. For example, the Chinese doesn't have white spaces between words.

For this reason, solving the problems described above is not enough in this language. In these cases, it is typically necessary to use external resources.

II. Classifications

The Classification task determines the type of each segment obtained in the segmentation task. In other words, it determines the classified output data structure where the inputs are segments. The result of this task is the classification of a set of segments as entities, which are elements of a given class potentially relevant for the extraction domain. The rule-based techniques used in the classification task are usually based on linguistic resources, such as lexicons and grammars. One of the most popular approaches to undertake classification is machine learning. Machine learning techniques used in this task are usually supervised, which means that an annotated corpus is needed.

III. Associations

The association task seeks to identify how the different entities found in the classification task are related. The systems that perform extraction of relationships are less common than the ones that perform the classification task. This happens due to the difficulty in achieving good results in this task. Many techniques in the association task are based on rules. The simplest approach uses patterns to extract a limited set of relationships. A more generic rule-based approach for association is based on syntactic analysis.

Often, the relationships that we want to extract are grammatical relationships. For example, a verb may indicate a relationship between two entities. The association task can also use machine learning techniques.

IV. Normalization and co-reference Resolution

Normalization and Co-reference resolution are the less generic tasks of the IE process since they use heuristics and rules that are specific to the data domain. The normalization task is required because some information types do not conform to a standard format.

This task is typically achieved through the use of conversion rules that produce a standard format previously chosen. Co-reference arises whenever the same real world entity is referred in different ways in a text fragment.

This problem may arise due to the use of

I. Different names describing the same entity (e.g., the entity Bill Gates" can be found in the text as William Gates"),

II. Classification expressions (e.g., a few years ago, Bill Gates" was referred as the world's richest man"),

III. Pronouns (e.g., in the sequence of sentences Bill Gates is the world's richest man. He was a founder of Microsoft", the pronoun He" refers to Bill Gates").

Rule-based approaches for co-reference usually take into account semantic information about entities.

A machine leaning approach for co reference resolution is based on clustering algorithms for grouping similar entities.

2.5. PREPROCESSING OF INPUT TEXTS

As described by Riloff [2] most of the text consists of unstructured, “raw” natural language texts. The relevant or necessary information can be distinguished by applying some linguistic properties of texts. In this phase the following linguistic components those are useful for IE will be described.

Tokenization*:* as it is defined by Siefkes and Siniakov [6], it is the process of splitting the text into sentences and tokens. It Start with a sequence of characters to identify the elementary parts of natural language such as words, punctuation marks and separators. As a sentence is one of the most important components in the natural language text for representation of interrelated information and for expressing a complete thought or event.

The resulting sequence of meaningful tokens is a base for further linguistic and any text processing task [6].

POS: Tokenization is similar to segmentation presented in POS: part of speech tagging, or simply tagging is the task of labeling (or tagging) each word in a sentence with its appropriate part-of –speech.

It is a technique for deciding whether each word is noun, verb, adjective, adverb, etc [2]. POS tagger has been applied to assign a single best POS to every word in corpus.

አበበ \NN\ **ትላንት**\AD\ **ሄደ**\VV Abebe went yesterday

In the above example, words in the sentence are tagged with appropriate lexical categories of noun, verb and verb respectively.

(Chunk) Parsing*:* While full sentence parsing is preferred by knowledge based systems, some statistical approaches rely on chunk parsing, shallow syntactic analysis of the sentence fragments performed on phrasal level. It is justified by the fact that the extracted information is often completely included in a noun, verb or prepositional phrase that builds the most relevant context for its recognition [2].

Co- reference resolution and named entity recognition are also included as a preprocessing component for IE [21].

2.6. LEARNING AND APPLICATION OF THE EXTRACTION MODEL

Learning phase is the back bone for designing IE system in any domain and language but it is not yet successfully as expected because of the morphology (the internal structure of the language) and the domain dependencies of IE.

Modern IE systems use a learning component to reduce the dependence on specific domains and to decrease the amount of resources provided by human. The three categories of approaches for IE use the learning methods are: Statistical approaches learn relevant classification features, probabilities, and state sequences, rule-based approaches learn a set of extraction rules and knowledge-based approaches acquire structures to augment and interpret their knowledge for extraction [6]

Most of the IE systems use the supervised learning approach to train the extraction model used about the domain specific information.

The statistical approach uses annotated training corpus which is divided in to two parts i.e. the training corpus and test corpus. The training corpus is used to training the model about the different annotation in the text and the test corpus is used to test the extraction model how much efficient it becomes after training [6].

2.7. POST PROCESSING OF OUTPUT

Once after the relevant information has been found by applying the extraction model on the given text the identified text fragments are assigned to the corresponding attributes of the target structure. They can be normalized according to the expected format.

Some identified facts may appear in text more than once and there might be violation of primary key and other properties of the database and all these things are handled at the post processing phase of IE [6].

2.8. RELATED NLP FIELDS TO INFORMATION EXTRACTION

2.8.1. INFORMATION RETRIEVAL (IR)

The meaning of the term IR can be very broad. Just getting a credit card out of your wallet so that you can type in the card number is a form of information retrieval. However, as an academic field of study IR might be defined as:

It is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) [22]. It is generally known that the number of resulting pages returned by the information retrieval system could be very large and not all information contained in a single page is interesting, which requires further refinement by the user.

As Eikvil [15] noted that, now the world has changed, and hundreds of millions of people engage in IR every day when they use a web search engine or search their email. IR is fast becoming the dominant form of information access, overtaking traditional data base style searching (the sort that is going on when a clerk says to you: "I'msorry, I can only look up your order if you can give me your Order ID"). IR can also cover other kinds of data and information problems beyond that specified in the core definition above.

2.8.2. TEXT SUMMARIZATION

It is the way of expressing a long text shortly by extracting the main points that are contain in original text or it is the process of creating abstract of one or more texts. It can also define as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s) [21].

According to Martins [14] the summary of text can be created from a single or multiple documents by applying different methodologies.

Although text summarization can reduce the size of text by filter out the most pertinent sentence from bulky text, the user involvement is mandatory to read the summary and filter out specific information she/he needs.

2.8.3. QUESTION ANSWERING SYSTEMS

IR has been researched mainly to help users in getting relevant documents from large collection of free-text documents. The way IR tackles the problem of document retrieval is based on the closeness of the document and the query submitted to the IR system. Strzalkowski and Harabagiu [27] notes, IR will not try to present answers to users explicitly. This was the critics of IR so that the need of IE came about.

The IE technique involves NLP tools for precisely indicating a correct text. There should be deep analysis of queries to understand the user's intention as well as deep analysis of the document to extract correct answers (sentences or passages). In the case of IR, a simple technique is sufficient to extract content-rich words from the query and applying stemming to make more uniformity of document retrieval that will be applied during indexing too.

Leidner and Burch [28] also acknowledge that Natural language question answering is profoundly different from IR or IE. IR systems locate relevant documents that relate to a query, but do not specify exactly where the answers for the users request are or where the specific information the user required is located. Even if the question and answering system extract answers from a document based on the input question from the user it still returns unstructured and small specific information from the document. The information that the Natural language question answering system returns can't be managed by the computer as it is more specific information to the question.

Therefore the question answering system is very helpful in extracting a specific answer for a specific question which reduces the time of the users greatly but it is limited in extracting all the information the user wants as it is much bound to the question formation in the language.

Most of question answering system encompasses the following things as the major components: Question Analysis, Document retrieval and Answer Extraction [26]. But the one local work that conducted by Imam [13], Amharic question answering system has mainly five components those are document pre-processing, question processing, document retrieval, sentence paragraph re-ranking, and answers selection modules. As the author noted that, considering document processing as a one components of the system could improves the performance of the system because Amharic is too specific in having different character representation with the same reading and writing style.

2.8.4. TEXT CATEGORIZATION

With the rapid growth of online data and information, text categorization has become one of the key techniques for processing and organizing text documents.

As it is defined by Sebastian [19] "is the task of automatically sorting a set of documents into categories from a predefined set". This task has several applications, including automated indexing of scientific articles according to predefined thesauri of technical terms, filling patents into patent directories, selective dissemination of information to information consumers, automated population of hierarchical catalogues of Web resources, identification of document genre, and authorship attribution.

The task of text categorization falls at the crossroads of IR and machine learning, has witnessed a booming interest in the last ten years from researchers and developers alike. [19]

2.9. INFORMATION EXTRACTION (IE) AND INFORMATION RETRIEVAL (IR)

IE is different from the more mature technology of IE. Rather than to extract information the objective of information retrieval is to select a relevant subset of document from a larger Collection based on a user query the user must the browse the returned documents to get the desired information.

The contrast between the aim of IE and IR system can be stated as follows: IR retrieve relevant document from collections, while IE extracts relevant information from documents. Hence, the two techniques are complimentary, and used in combination they can be provide powerful tool for text processing [20].

Cunningham [22] is also tried to differentiate IR and IE as follows:

- An IR system finds relevant texts and presents them to the user;

- An IE application analyses texts and presents only the specific information from them that the user is interested in. For example, a user of an IR system wanting information on trade group formations in agricultural commodities markets would enter a list of relevant words and receive in return a set of documents (e.g. newspaper articles) which contain likely matches. The user would then read the documents and extract the requisite information themselves. They might then enter the information in a spreadsheet and produce a chart for a report or presentation. In contrast, an IE system would automatically populate the spreadsheet directly with the names of relevant companies and their groupings.

2.10. EVALUATION OF INFORMATION EXTRACTION

IE systems may take quite diverging approaches in solving the problems at hand. A fair comparison of the results is often not directly possible.

We need a comparison method in order to decide which approach works better under given circumstance. That will enable us to compare IE results.

Additionally the performance of single components of IE needs to be evaluated. Johannes noted that [1] most often the used evaluation method is statistical evaluation. The results are compared against the correct solution to the problem. However, we have to keep in mind that sometimes there is not only one correct solution. In some cases even human experts disagree on which information exactly to extract.

If we assume that a correct solution is available, it consists of a data tuple for each relevant text in the test corpus, containing the correct extraction set of information. An extraction result for a text is correct when the correct data tuple matches exactly the template the IE system a result is incorrect when the IE system extracts information that does not match, or that is irrelevant to the correct solution.

An exact match of one data tuple is called a true positive (TP). A text document considered irrelevant in both the correct solution and the IE system output is called a true negative (TN). If the IE system discards a text document which is not discarded in the correct solution, we call it a false negative (FN). If the correct solution rejects a document but the IE system extracts a data tuple from it, it is a false positive (FP).

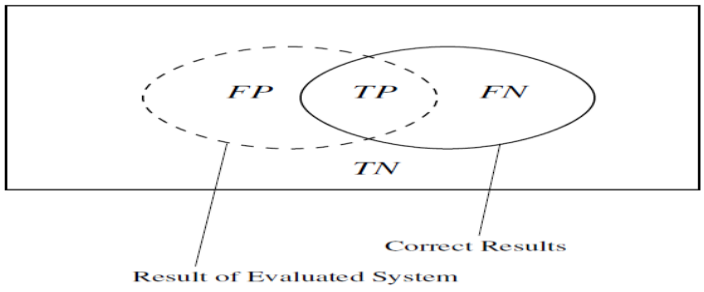

Figure 2.1 TP, TN, FP and FN as a set

As fig 2.1 depicts, the result is correct when both sets, the evaluated results and the correct results, are the same and FP = FN = $\emptyset$. we need to compute more expressive values in order to get useful evaluation attributes from these statistical values.

The precision π, given in equation $\pi = \frac{TP}{TP + FP}$ describes how many of the results the IE system generates are correct. The recall ρ, given in equation $\rho = \frac{TP}{TP + FN}$, describes how many of the correct results the IE system is able to find. Obviously precision and recall range in the interval $0 \leq \pi, \rho \leq 1$. Precision and recall each describe one quality of the IE system's performance.

Systems have to maximize both. Accordingly, full evaluation of the system is only expressed by both values.

In other words, if the precision is very high and the recall rather low, the result may be worse than an average value for both precision and recall. If we want to compare systems directly, we need one single value representing the quality of the IE system performance.

2.11. RELATED WORKS

Many Scholars tried to develop IE system using different approaches on different domains and languages. Among these, we are going to review some of the works those are more relevant to this research.

INFORMATION EXTRACTION FOR E-JOB MARKETPLACE

The work in [20] presents IE for E-Job Marketplace. The main motive of this work was applying IE technique to locate and extract specific and important data about from the available job postings on the Internet.

The slots (attribute) for job that they used are: Job Title, Industry, Level of education, Work Experience (in years), and Language Skill.

In order to extract relevant information from a text they construct a rule. As they noted, the important element for IE is a set of extraction rules, which identifies relevant information to be extracted. Every domain has different set of rules; which can be constructed manually or generated automatically.

This system adopted knowledge engineering approach as a baseline. Based on their examination of 25 job postings, they manually crafted 14 extraction rules for 5 slots and the system can perform well, the experimental result showed 88.8% of precision and 37.5% of recall.

According to the authors merging job information from Websites that use free or semi-structured document is very useful for developing intelligent e-job marketplace that supports high precision IR, job recommendation and job summarization.

INFORMATION EXTRACTION FROM AMHARIC TEXT

In 2010 Tsedalu [17] proposed model for Amharic Text IE based on which he designed an IE system which is called ATIE (Amharic Text Information Extraction) system. Users who want to use the available unstructured information for different purposes have to read all the relevant texts that are related to their need and have to manually extract the information they want from the abundant unstructured text which takes very long time and which in turn highly minimizes the users efficiency in decision making. The IE system developed for English or any other language and for some specific domain cannot work for other languages of the same domain, this and other problems are initiated the researcher to do a research on IE for Amharic text. Document Preprocessing, text Categorization Component, Learning and Extraction Component, and Post Processing are the four components of the model and it is developed using the open source machine learning algorithms in Weka and java programming language.

He conducted an experiment specifically on economy news category obtained from Ethiopian News agency. He considered Infrastructure and Investment as subcategories for text categorization and he used six predefined attributes are for IE. His experimental result shows that classifier algorithms can be used and perform as that of the other algorithms for IE.

Among the three different classifier algorithms that he used for experimentation the first algorithm performs better than other on both text categorization and IE.

The token category feature plays a crucial role in increasing the performance of the classifier when compared to that of Part of speech. Since his Part of speech performance is 80% only but he claimed that, using another POS which perform better in predicting the POS tag might increase the performance of the prediction.

Finally he showed that the confusion matrix in each of the four scenarios for IE is small. This work is limited only in extracting numeric and name data from the Amharic news text and extraction of relationship between entities is out of the scope of this research work. News texts that are about a single issue are only considered. Generally this system cannot fully handle the concerns that are viewed in data extraction from AVAT.

INFORMATION EXTRACTION FROM ENGLISH TEXT

Most of the research works in the area of IE are conducted in different domain on the English language. Among these the research that done by Rosendfeld et al. [22] is the one, that presents a hybrid approach to IE. The authors present a hybrid knowledge based and statistical machine learning approach to extract entities and relations at the sentence level. They present a hybrid entities and relation extraction system, which combines the power of knowledge based and statistical machine learning approach. In this approach, the rules for extraction are written manually, while the probabilities of the extracted texts being part of the database slot are trained from an annotated corpus. This approach allows the knowledge engineer to write very simple and naive rules, while retaining their power thus greatly reducing the required labor.

In addition the size of the training data is considerably smaller than the size of the training data needed for pure machine learning system (for achieving comparable accuracy results). The Authors use DIAL, which is based on a general purpose rule language for developing knowledge for extraction and statistical HMM for machine learning approach which is used to train the system.

Effectiveness of their approach is tested by using three different corpora MUC-7, ACE -2 and an industry corpus. On the MUC-7 corpus their hybrid approach called TEG outperforms pure HMM model and DIAL rule based system for named entity recognition. On the ACE -2 corpus they tested the relationship and in this case as well the hybrid approach they present performs better than HMM and Markovian SCFG.

They conclude that a small hand crafted rules when combined with machine learning method will increase the performance of machine learning.

INFORMATION EXTRACTION FROM CHINESE TEXT

A work in [25] presents an IE from Chinese free text. The authors present an approach which combines Automatic learning algorithm of pattern rule and employment of heuristic information for Chinese free text. The authors present the different tasks they use to extract information from Chinese free text.

Input Document Preprocessing: this phase contains different subtasks to make the Chinese free text ready for the next phase. At first the input document is broken down into sentences. Then the sentences are segmented into words by looking up the dictionary because Chinese sentence is composed of characters without any natural delimiters such as space between words. After the sentence segmentation the named entity recognition is done to identify the place, person, and organization names.

Syntactic Analysis: uncategorized words during preprocessing are assigned in to the likely category based on the HMM model probability approach.

VP and NP Recognition: verb and noun signify the important meaning of sentence at most part. This process module uses syntactic patterns to identify small syntactic units through rule primarily, such as basic noun groups (NG e.g. position group, organization group…), which are nouns with their left modifiers, and verb chains or verb groups (VG), which consist of a head
Verb preceded by modals or adverbials.

Pronominal Anaphora Resolutions: It is very difficult to understand the extracted contents in some instances that pronouns are contained without their antecedents. So, it is important to resolve pronominal anaphora.

Pronominal anaphora resolution is to track references to a frame topic across sentences.

For each finding, the probability is estimated that it co-refers to each previously mentioned finding based on semantic features and dictionary cues. Their pronominal anaphora resolution work uses rule and statistical methods.

Use of Heuristic Information: Also extraction pattern can play an important role in IE system, some instances test are not accorded with existing pattern. In order to solve this problem, more complicated pattern is needed to face to complex situations. This will bring out more difficult to system and make conflict in future. They employ heuristic information to reduce this complexity.

The heuristic information is used to exploit the context structure of the source. It will be simplify the structure of the text so that the pattern rules can be used to extract the information. Then the extraction rules are used to extract the text fragment and fill the template slot after the above different processing is done on the Chinese free text.

Their approach is tested on 50 articles they get from 1 China import official alteration to extract information that will fill the following four slots person name, organization, old position and new position. They test the system by applying two methods; method 1 is based on pattern matching without heuristic information and method 2 is based on pattern matching with heuristic information. And the result of the experiment was the following

Slot & result		Method 1	Method 2
Person name	Recall	64.1%	78.8%
	precision	89.2%	87.2%
Organization	Recall	62.3%	76.5%
	Precision	92.1%	89.3%
Old position	Recall	64.5%	77.4%
	Precision	86.3%	84.6%
New position	Recall	68.3%	83.3%
	precision	84.5%	81.3%

The authors conclude that the use of heuristic information in addition to pattern learning will increase the efficiency of IE specially the Recall.

CHAPTER THREE
THE AMHARIC WRITING SYSTEM

3.1. INTRODUCTION

Amharic was the national language of Ethiopia until 1983 E.C. Currently it is the official working language of the Federal Democratic Republic of Ethiopia and thus has official status nationwide and the official or working language of several of the states/regions within the federal system, including Amhara and the multi-ethnic Southern Nations, Nationalities and Peoples region. It was also for a long period the principal literal language and medium of instruction in primary and secondary schools in the country, while higher education is carried out in English [30]. Currently, different Mass Medias like radio, television broadcasts and the press are also using it for disseminating information to the public.

[Outside Ethiopia, Amharic is the language of millions of emigrants (notably in Egypt, US, Israel, and Sweden), and is spoken in Eritrea. It is written using a writing system called fidel, adapted from the one used by Ge'ez language [35].

Geez has been a language of literature in Ethiopia up to recent time and is now used for the liturgy of the Ethiopian Orthodox Church. Written Geez can be traced back to at least the 4^{th} century A.D. The first versions of the Geez script included only consonants while the characters in the later versions represent consonant-vowel (CV) phoneme pairs [30].

As a result of its wide application, currently, large Amharic documents are compiled in electronic forms. Due to this, the amount of electronic Amharic information is increases from time to time, thus, it is mandatory to perform a task of NLP and utilize the knowledge that confined in natural languages.

For the purpose of this research since Amharic texts are considered, it is important to investigate the characteristics and morphology of the language. Hence, under this section, the grammatical structure of the Amharic language those are believed to be pertinent to the current research will be reviewed.

3.2. AMHARIC CHARACTER REPRESENTATION AND WRITING SYSTEM

Amharic has borrowed most of its characters from Geez and thus the Amharic writing uses characters created by a consonant-vowel fusion. Sven vowels are used in Amharic each of which comes in seven different forms (orders) reflecting the seven vowel sounds (ኧ ኡ ኢ ኣ ኤ እ ኦ). That is each of the 33 Amharic characters has seven forms representing a consonant and a vowel at the same time which makes the Amharic script syllabic. The first order is the basic form and there are 33 basic forms giving 231 characters [31].

As examples, the symbolic representations of the seven forms of the Amharic characters በ(be) and ገ(ge) as shown in Table 3.1.

Consonant	1st order	2nd order	3rd order	4th order	5th order	6th order	7th order	
በ	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ	the seven forms of በ
	ብኧ	ብኡ	ብኢ	ብኣ	ብኤ	ብ	Bå	Consonant-vowel representation
	Bä	bu	bi	Ba	be	B	bo	Represented sound
ገ	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጐ	the seven forms of ገ
	ግኧ	ግኡ	ግኢ	ግኣ	ግኤ	ግ	ግኦ	consonant-vowel representation
	Gä	gu	Gi	Ga	Ge	G	go	Represented sound

Table 3.1 Seven forms of Amharic Characters Consonant

3.3. AMHARIC PUNCTUATION MARKS AND NUMERALS

I. Punctuation

In Amharic, there are different punctuation marks used for different purposes. Among this, we discussed some of the punctuations those which are important to this research.

- "ሁለትነጥብ:" (two dots) is used to separate two words, this punctuation was commonly used in the old scripture but these days the two dots are replaced with whitespace.
- "አራትነጥብ።"(four dots) is always take the place of the end of the sentence
- ነጠላሰረዝ(፣or ፥) is used to separate lists or ideas; it acts as just like a comma in English. Connection

II. Numerals

The numbers in Amharic can be represented in three way; Arabic number system, the symbols of the Ethiopic number system, and using words and symbols of the Arabic number system.

Arabic	Ethiopic	Alphanumeric	Arabic	Ethiopic	Alphanumeric
1	፩	አንድ	20	፳	ሃያ
2	፪	ሁለት	30	፴	ሰላሳ
3	፫	ሦስት	40	፵	አርባ
4	፬	አራት	50	፶	አምሳ/ሀምሳ
5	፭	አምስት	60	፷	ስልሳ/ስድሳ
6	፮	ስድስት	70	፸	ሰባ
7	፯	ሰባት	80	፹	ሰማኒያ
8	፰	ስምንት	90	፺	ዘጠና
9	፱	ዘጠኝ	100	፻	መቶ
10	፲	አስር	1000	፼	ሺ/ሽህ

Table 3.2 Number Representations in Amharic

In Amharic, fractions and ordinals also have their own way of representation. Table 3.2 shows fraction and ordinal representations in Amharic. As numbers are one of the information that is extracted in this research work its representation in letters in Amharic text is important and it is presented in the following table.

Fraction	Amharic representation	Ordinals	Representation
½	ግማሽ	1st	አንደኛ
1/3	ሲሶ	2nd	ሁለተኛ
¼	ሩብ	3rd	ሦስተኛ
2/3	ሁለትሲሶ/ ሁለትሦስተኛ	4th	አራተኛ
¾	ሶስት-አራተኛ	.	.
1/10	አስራት	.	.
2.X	ሁለትነጥብ	10th	አስረኛ

Table 3.3 Amharic Fraction and Ordinal Representation (Adopted from Tsedalu [6])

In Amharic dates can be written by using symbols in Arabic number system like 19/07/2005 or using Ethiopic numeral representation and alpha-numeric representations like የካቲት19/2005.

3.4. CHARACTERISTICS OF THE AMHARIC WRITING SYSTEM

As it was discussed in many literatures, the Amharic writing system has many features, which may cause some problem from the perspective of computation [2, 3].The next section deals some of them.

I. Consonants with the same sound: At the time of borrowing its script from Geez, Amharic did not select consonants which are only important to its writing system. As a result, in Amharic writing system, there has been found different symbols with the same pronunciation and meaning (i.e., in Geez those symbols are different in meaning as well as in spelling, which is not the case for Amharic) and they have been used interchangeably [2, 3].

As Getachew [31] noted, for the case of Amharic there is no defined rule that differentiates their proper usage. In Amharic, these consonants with the same sound falls into two categories: (1) the first and the fourth order alphabets of the same base form having the same sound and (2) different alphabets with the same sound. For the first case, for instance, it is not clear whether one should write "ሃይማኖት" (`religion) and"ሀይማኖት" since both "ሃ" and "ሀ" have the same sound. Those alphabets that exhibit such characteristics are listed in table 3.4.

1st order	**4th order**
ሀ(hä)	ሃ(hä)
ሐ(hä)	ሓ(hä)
ኀ(hä)	ኃ(hä)
አ(ä)	ኣ(ä)
ዐ(ä)	ዓ(ä)

Table 3.4 different forms of the base alphabet with the same sound

Similarly, table 3.5 shows lists of different alphabets that have the same meaning and sound. Here, not only the base forms listed have the same sound but also all the corresponding orders (6orders) of them have the same sound too. For example, writing "ሰማይ" and "ሠማይ" to mean "the sky" does not make difference in meaning even though "ሰ" and "ሠ" are used interchangeably. The same holds true for "አይን”" (eye) and " ዐይን”" although "አ" and "ዐ" are two different alphabets with the same sound.

Alphabet	**Other alphabet with the same sound**
ሀ(hä)	**ሐኀ**
ሰ(sä)	**ሠ**
አ(ä)	**ዐ**
ጸ(tsä)	**ፀ**

Table 3.5 Different alphabets having the same sound

Moreover, a complex case comes when the same word appears to be in many forms (more than two forms) by using interchangeably these alphabets having the same sound. We can take "ገብረስላሴ", "ገብረሥላሴ", "ገብረስላሤ" and " ገብረሥላሤ" as a good example, which refers to the name of a person (Gebresilase). As all the above discussion indicates, there arises some confusion and inconsistencies in Amharic alphabet and as a result these redundant consonants add their contribution to make the vocabulary to be large.

Also it is obvious that, spelling variations of a word would unnecessarily increase the number of words representing a document which could reduce the efficiency and accuracy of the system. During the pre-processing stage of Amharic documents for this research, the different forms of a character that have the same sound are changed to one common form.

II. Different forms of writing compound Nouns: In Amharic writing system, there also exist different ways of writing compound words without affecting their meaning (Bender and Ferguson, 1964) as cited by Bizuneh [33]. That means, at one time the compound noun can be written as two separate words and at another time as single word. For instance, it makes no difference in meaning at all while writing the compound word ወጥቤት" as one word "ወጥ ቤት" which is to mean that "Kitchen". Additional examples of such Nouns are mentioned in table 4.5.

Compound Noun as two separate words	**Compound Noun as single word**	**Its meaning in English**
ማዕድቤት	ማዕድቤት	Dining room
ብርድልብስ	ብርድልብስ	Blanket
ብረትድስት	ብረትድስት	Cooking pot (metallic)
ቤተመቅደስ	ቤተመቅደስ	Temple

Table 3.6 some examples of writing compound nouns in different ways (adapted from Bizuneh [5])

III. **Different ways of writing the same word:** This is a problem that exists when the language has some words having different forms of writing system. When we write English language word by Amharic characters, the same word may have different spelling due to the reason of Pronunciation [31].That means different peoples may pronounce the same foreign language word in various ways, therefore, this makes spelling variation in Amharic writing system. The cause of the difference in the Amharic spellings of these foreign language words seems to be the difference in the pronunciations of these words.

In many Amharic vacancies announcement texts some of the words are adapted from English languages, at this time, the writers write those words in different spellings, due to the reason of what we discussed in above.

For example, the word ሚቲዎሮሎጂ(Meteorology) is found to have 14 different Amharic spellings in the source data. Table 4.6 shows examples of spelling variation in the writing of foreign words in Amharic.

Foreign Word	**Equivalent Words in Amharic usage**
Meteorology	ሚትሪዮሎጂ፣ሚትዎሮሎጂ፣ሚትዮሮሎጂ፣ሚቲዎሮሎጂ፣ሚቲዎሮሎጅ፣ ሚቲዎሮሊጂ፣ሜትዎሮሎጂ፣ሜትሮዎሎጂ፣ሜትሮዎሎጅ፣ሜትሪዎሎጂ፣ ሜትሮሎጂ፣ሜትርዎሎጂ፣ሜቲዎሮሎጂ፣ሜቴዎሮሎጂ
Telecommunication	ቴሌኮሚዩኒኬሽን፣ቴሌኮሙኒኬሽን፣ቴሌኮሚኒኬሽን፣
Million	ሚሊዮን፣ሚልዮን፣ሚሊየን

Table 3.7 some examples of word variation due to translation

IV. **Different forms of writing Abbreviation:** In Amharic, it is also found that there is no consistency while spelling abbreviations. For instance, the phrase "ዓመተምህረት" can be abbreviated as "ዓ.ም", "ዓም" and "ዓ/ም". Similarly, the use of hyphen is also not consistent. The same word "ዓመተምህረት" can also be written as "ዓመተ-ምህረት". Hence there should be a mechanism to handle these problems while representing Amharic documents.

3.5. THE MORPHOLOGY OF AMHARIC

As oxford dictionary define Morphology in this context “it indicates how the words of a given language are formed”. Every natural language has its own morphological structure and defines rules for the different components of the language.

As we know Amharic is one of the morphologically rich languages. According to Bender [32], like other Semitic languages such as Arabic, Amharic exhibits a root-pattern morphological phenomenon. A root is a set of consonants (called radicals) which has a basic ’lexical’ meaning.

A pattern consists of a set of vowels which are inserted (intercalated) among the consonants of a root to form a stem. The pattern is combined with a particular prefix or suffix to create a single grammatical form.

For example, the Amharic root sbr means ’break’, when we intercalate the pattern ¨a ¨a and attach the suffix ¨a we get s¨abb¨ar¨a ’he broke’ which is the first form of a verb (3rd person masculine singular in past tense as in other Semitic languages) [32].In addition to this non-concatinative morphological feature, Amharic uses different affixes to create inflectional and derivational word forms.

As it is discussed in [35] some adverbs can be derived from adjectives. Nouns are derived from other basic nouns, adjectives, stems, roots, and the infinitive form of a verb by affixation and intercalation. For example, from the noun lI˘g˘ ’child’ another noun lI˘gn¨at ’childhood’; from the adjective d¨ag ’generous’ the noun d¨agn¨at ’generosity’; from the stem sInIf, the noun sInIfna ’laziness’; from infinitive verb m¨asIb¨ar ’to break’ the noun m¨asIb¨ariya ’an instrument used for breaking’ can be derived. Case, number, definiteness, and gender marker affixes inflect nouns. Adjectives are derived from nouns, stems or verbal roots by adding a prefix or a suffix.

For example, itis possible to derive dIngayama ’stony’ from the noun dIngay ’stone’; zIngu ’forgetful’ from the stem zIng; s¨an¨af ’lazy’ from the root snf by suffixation and intercalation.

3.6. GRAMMATICAL STRUCTURE OF AMHARIC

3.6.1 WORD CATEGORIZATION IN AMHARIC

The words in Amharic are categorized under five basic categories by Yimam [33] that uses the morphology and position of the word in Amharic sentence as criteria. These five categories are ስም(noun), ግስ(verb), ቅፅል (adjective), ተውሳከግስ(Adverb) and መስተዋድድ (preposition).

I. **Noun:** a word will be categorized as a noun, if it can be pluralized by adding the suffix ኦች and used as nominating something like person and animal. It is used as a subject in a sentence.

   Pronouns, which were considered as independent category in the previous works by the linguistics professionals is categorized under nouns after considering the unique nature of the language as the earlier linguists just adopt the English language structure for Amharic language.

   The following are some of the pronouns in Amharic ይህ, ያ, እሱ, እስዋ, እኔ, አንተ, አንች...;quantitative specifiers, which includes አንድ, አንዳንድ, ብዙ, ጥቂት, በጣም...; and possession specifiers such as የእኔ, የአንተ, የእሱ.

II. **Verb:** any word which can be placed at the end of a sentence and which can accept suffixes as /ህ/,/ሁ/,/ሽ/, etc. which is used to indicate masculine, feminine, and plurality is classified as a verb**.**

   For example in "አበበ አንበሳ ገደለ" "ገደለ" is a verb since it appears at the end of the sentence.

III. **Adjective**: is a word that comes before a noun and add some kind of qualification to the noun. But every word that comes before a noun is not an adjective. For it to be an adjective it should also satisfy the condition when the word "በጣም" is added to it, it should be meaningful.

For example “ትልቅ በግ” in this example “ትልቅ” is an adjective to check it really is an adjective adding the word “በጣም” before the adjective if it is meaningful it is an adjective if not is isn’t an adjective. In this case it is meaningful and “ትልቅ” is an adjective.

IV. Adverb: a word that qualifies the verb by adding extra idea from time, place and situations point of view. The following are adverbs in Amharic ትናንት, ገና, ዛሬ, ቶሎ, ምንኛ, ክፉኛ, እንደገናand ግምኛ.

V. Preposition: a word that doesn’t take any kind of suffix and prefix, that can’t be used to create other words and which doesn’t have meaning by itself but can represent different adverbial roles when used with nouns. The different propositions include ከ፣ለ፣ወደ፣ስለ፣እንደ...፣ወዘተ

3.7. SENTENCES IN AMHARIC

As discussed in [17] a sentence in Amharic can be a statement which is used to declare, explain, or discuss an issue. The combination of phrases to create another phrase that can express a full idea on something is a sentence.
When Amharic sentence is viewed from grammatical structure point of view it is a combination of noun phrase and verb phrase. The noun phrase comes first and then the verb phrase. Based on the number of phrases they contain sentences in Amharic are categorized under two basic categories simple sentence and complex sentence. Simple sentence only contains a single verb while complex sentence is constructed by combining more than one noun phrases and verb phrases.

CHAPTER FOUR
DESIGN AND IMPLEMENTATION OF AVATIES

4.1. INTRODUCTION

This chapter can be considered as the central part of this study. Based on the assumptions and approaches discussed in chapter two and the syntactic property of Amharic reviewed in chapter three, this chapter discusses the propose IE for Amharic vacancy announcement text Model (Figure 4.1), the main components of the model along with their subcomponents and the interaction between the main components.

Afterward, the Amharic text Information extraction prototype system that is developed based on the proposed model, the resources and algorithms used in the three main components of the system and how the candidate texts for extraction are identified from unseen AVAT are also discussed.

4.2. PROPOSED MODEL

The model, which is used to design IE for AVAT is developed by the researcher. The main reason that required for designing new model is: first the unavailability of any effective IE model for Amharic language and specially for this domain, even though there is a one model which was designed by Tsedalu[17], it is not fully applicable as intended for this domain, because the researcher has designed the model by making compatible only for his targets, that mean his intention was only extracting the nominal and numerical data from infrastructural news type. likewise, other models, those are developed for other language most of them are used machine learning approach, they employed different natural language processing tools such Sentence Parser, Part of Speech tagger (POS), Named Entity Recognizer (NER), Co-reference Resolution and others. Even though some of the NLP systems for Amharic language have been conducted by some researchers, as we have said in section 1.6, they are not fully available and might not easily applicable.

Due to this and other reasons the researcher decided to develop a new model for this study rather than simply adopt the one which is already designed for other language and domain.

Johannes [1] acknowledged that, every IE system has three basic components which are the linguistic preprocessing, learning and extraction and post processing regardless of the approach, language and domain on which the IE system is developed for. In addition to these three components other subcomponents are also included in each of the main components.

The model which is designed in this study is also has the three major components and these three main components also contain different subcomponents which are language specific and general subcomponents that are required in IE.

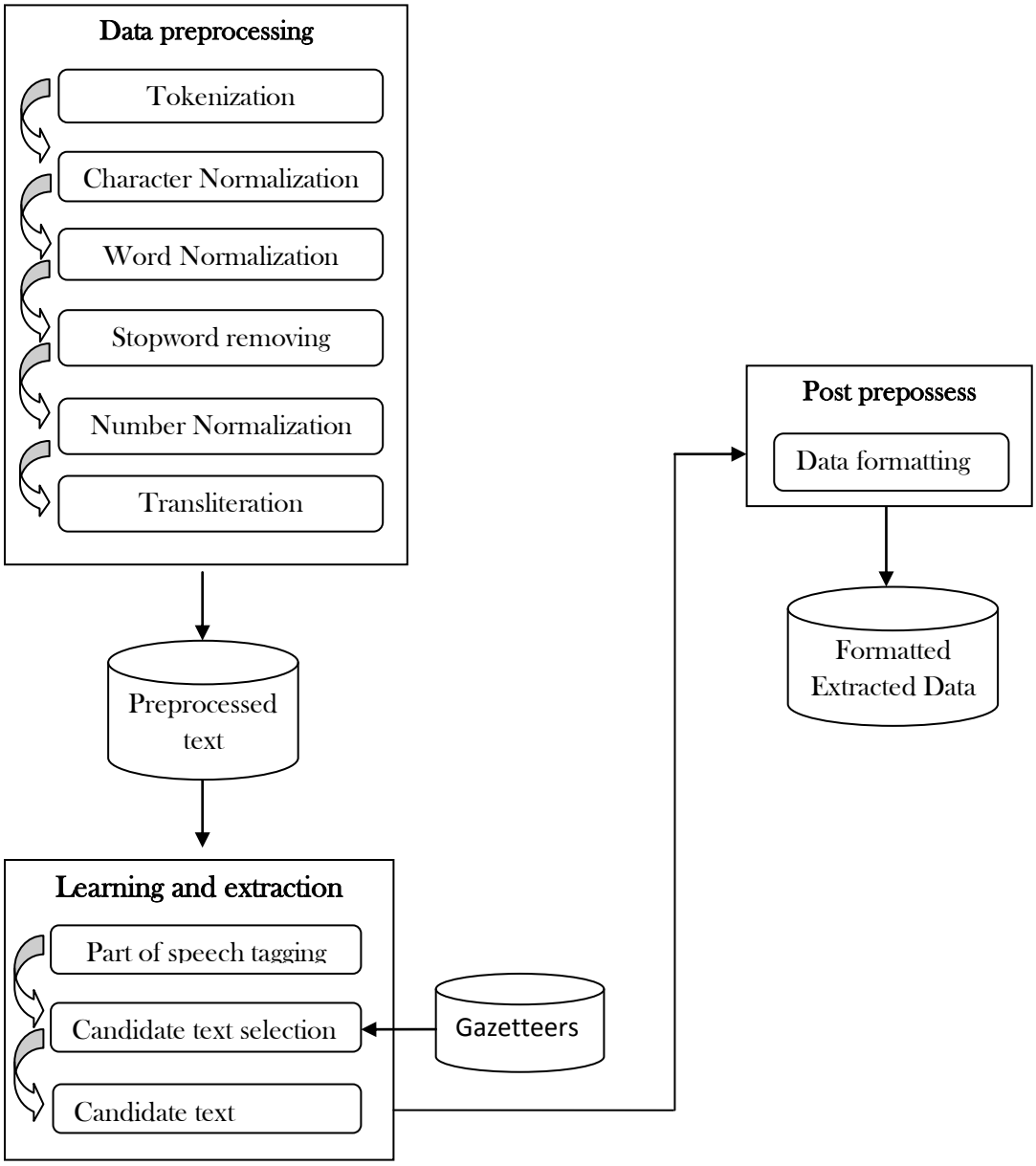

Fig 4.1 ***AVATIES model***

In the following subsections the three components of the model will be discussed.

DATA PREPROCESSING

Preprocessing is an important part of all text processing. In the preprocessing stage file formats, character sets, and variant forms can be converted, so that all text, regardless of its source, is in the same format. In later stages all further processing can then be consistently applied to all of the data. Preprocessing must ensure that the source text be presented to NLP in a form usable for it. For example, NLP programs usually need their input to be tokenized, i.e. text elements usually word forms or sentences are identified and placed on separate lines of the input.

In the data preprocessing stage a language specific issue such as tokenization, normalization, stop word removal and transliteration are addressed in this study.

I. Tokenization

It is generally known that; tokenization is an important step in NLP particularly for information extraction system. As we know there is no single right way to do tokenization. The right algorithm depends on the application.

In this study, words are taken as tokens. All punctuation marks (except "/"), control characters, and special characters are removed from a text before the data is transferred for further process. / (ህዝባር) the Amharic slash has its own role during text normalization, due to this; it would not be removed during this process.

The tokenizer, which is adopted for the purpose of this study is used the following punctuations more prominently, such as ። (አራት ነጥብ) the Amharic full stop and ፣ (ነጠላ ሰረዝ) the Amharic comma for tokenization process, because they are the most commonly used punctuation marks in the AVAT.

The። (አራት ነጥብ) the Amharic full stop is used for identifying the sentence demarcation and (ነጠላ ሰረዝ) the Amharic comma is used to separate different text segments which mostly are related.

When these punctuations are found in a text a single space would be added between the word and punctuations by the system. The tokenizer then tokenizes all the text segments which have space between each other as independent token. The algorithm which we used in this study for tokenization is as shown below.

```
Read raw corpus
While (sentences! =null)
String PUNCT = "$# {} []:!\"%&/?-*()_+,\t\n ፡፤
While (string! =PUNCT) do
        Concatenate characters in the string
Break the string
```

Figure 4.2: Tokenizer Algorithm

II. Character normalization:

It is generally known that, in Amhric writing system different characters with the same sound are available. These different symbols must be considered as equivalent because they do not cause changes in meaning in IE system. Though from the lingustics view this character variation might be have meaning, they need to be normalized when developing IE system for Amharic language, because spelling variations of a word would unnecessarily increase the number of words representing a document, which could reduce the efficiency and accuracy of the system.

The letters such as ሀ, ኀ, ሃ, ኻ, ሐ, ኃ and ሐ, ዐ and አ, ሠ and ሰ, ፀ and ጸ, are the characters with the same meaning and pronunciation but different symbol. this character variation also exists in most of AVAT.

These characters should be normalized to a single characters like ሠ to ሰ and ኀ, and ሐ to ሀ and ፀ to ጸ as well as their orders (ሠ, ሡ, ሢ, etc. to ሰ, ሱ, ሲ, etc.) consequently.

In addition to this, we further investigated and found that some other orders of the letters should also be normalized. For example ሀ, ኀ, ሐ, ሃ, ኃ, and ሓ should be normalized to ሀ. Using these characters as they existed in the document will be considered as different from one another in the training and testing of IE system.

```
Read raw corpus
Read list of similar charactors
String = each token in raw corpus
Char = each charactor in list of similar charactors
For String
    For char
        If string == char
            Normalize string
        Else continue
        End if
    End for
End for
```

Figure 4.3 Charactor Normalizer Algorithm

Characters with the same sound	**Normalized to**
ኅ፣ኍ፣ኂ፣ኃ፣ኄ፣ኅ፣ኆ ሐ፣ሑ፣ሒ፣ሓ፣ሔ፣ሕ፣ሖ ኸ፣ኹ፣ኺ፣ኻ፣ኼ፣ኽ፣ኾ	ሀ፣ሁ፣ሂ፣ሃ፣ሄ፣ህ፣ሆ
ሠ፣ሡ፣ሢ፣ሣ፣ሤ፣ሥ፣ሦ	ሰ፣ሱ፣ሲ፣ሳ፣ሴ፣ስ፣ሶ
ፀ፣ፁ፣ፂ፣ፃ፣ፄ፣ፅ፣ፆ	ጸ፣ጹ፣ጺ፣ጻ፣ጼ፣ጽ፣ጾ
ዐ፣ዑ፣ዒ፣ዓ፣ዔ፣ዕ፣ዖ	አ፣ኡ፣ኢ፣ኣ፣ኤ፣እ፣ኦ

Table 4.1 Amharic Characters which have different symbols but similar sound

III. Word normalization

In Amharic language the one foreign could write in different Amharic writing system due to the reason of pronunciation. That means different peoples may pronounce one foreign language word in various way, therefore, this makes spelling variation in Amharic writing system.

In AVAT also some foreign language those adapted from English languages word is identified specially on job position and organization name of the vacancy announcement, as a result of this, the tendency for word variation in AVAT is became high.

For example, the word ሜትሮሎጂ (Meteorology) is found to have 14 different Amharic spellings in the source data. Table 4.1 shows examples of spelling variation in the writing of foreign words in Amharic.

In designing any IE system such word should be normalized to the one word which can represent all synonyms words adequately. Some of the common words those found in AVAT are listed in table 4.2 with their normalized word.

Moreover, in Amharic writing system different words or phrases can be write using different abbreviation which can represent the word in shortest forms. For example መስሪያ ቤት can be reprerseented by መ/ቤት and ሃላፊነቱ የተወሰነ የግል ማህበር can be reprerseented by ሃ.የተ.የግ.ማህበር or ሃ/የተ/የግ/ማህበር or ሃላ/የተ/የግ/ማህበር etc. So such abbreviations those found in our data set must be normalized or converted in to formal Amharic writing way, because they have an impact on extraction of candidate texts, since our proposed system approach is a rule-based, which used context information of the candidate texts.

```
Read raw corpus
Read list of similar words
String = each token in raw corpus
simword = each word in list of similar words
    For String
        For simword
            If string == simword
                Normalize string
        Else continue
        End if
    End for
End for
```

Figure 4.4 word Normalizer Algorithem

Equivalent Words in Amharic usage	**Normalized to**
ሚትሪዮሎጂ፣ሚትዎሮሎጂ፣ሚትዮሮሎጂ፣ ሚቲዎሮሎጂ፣ሚቲዎሮሎጅ፣ሚቲዎሮሊጂ፣ ሜትዎሮሎጂ፣ሜትሮዎሎጂ፣ሜትሮዎሎጅ፣ ሜትሪዎሎጂ፣ሜትሮሎጂ፣ሜትርዎሎጂ፣ሜቲዎሮሎጂ፣ ሜቴዎሮሎጂ	ሜትሮሎጂ
ቴሌኮሚዩኒኬሽን፣ ቴሌኮሙኒኪሽን፣ ቴሌኮሚኒኬሽን፣	ቴሌኮሚኒኬሽን
ሚሊዮን፣ ሚልዮን፣ ሚሊየን	ሚሊዮን
ቴሌቭዥን፣ ቴሌቪዢን	ቴሌቪዢን
ሾፌር	ሹፌር
አ/አ	አዲስ አበባ
አ/ማህበር, አ.ማህበር, አ.ማ	አክስዮን ማህበር
ት/ቤት፣ ት/ቤ ት.ቤ	ትምህርት ቤት
ሃ.የተ.የግ.ማህበር, ሃ/የተ/የግ/ማህበር, ሃ.የተ/የግ/ማ, ሃ.የተ.የግ.ኩባንያ,ሃላ.የተ.የግ.ማህበር	ሃላፊነቱ የተውሰነ የግል ማህበር

ጽ/ቤት	ጽፈት ቤት
ስ/አስኪያጅ	ስራ አስኪያጅ
የት/ደረጃ	የትምህርት ደረጃ
ተ/ችሎታ	ተፈላጊ ችሎታ
ያለው/ት	ያለው ያላት
የተመረቀ/ች	የተመረቀ የተመረቀች
ስ.ቁ , ስ/ቁ	ስልክ ቁጥር
ሲ/ኦፐሬተር	ሲኒየር ኦፐሬተር
ሲ/አካውንታንት	ሲኒየር አካውንታንት
ሲ/መሃንዲስ	ሲኒየር መሃንዲስ

Table4.2. Amharic words those are normalized

IV. Number normalization

In amharic vacanncy announcment text there are diffrent entites which are represented by number, among these salary is the one.

There is no any standard for writing numbers in Amharic; someone may write by using only Arabic numerals, others may write by mixing both Ethiopic numerals with Arabic numerals. For example in most of the AVAT the salary 8000 birr is written as “8 ሺህ 000 or 8,000.00 or ስምንት ሺህ birr. This representation in the AVAT causes problem during information extraction.

The number Normalizer changes all above listed of number representation in to their equivalent number representation. For example “8 ሺህ 000” or “8,000.00” or “ስምንት ሺህ” birr will be normalized in to 8000

V. Transliteration

The dataset used in this research is published using the Ethiopic script with a variety of fonts. For the purpose of POS tagger that employed in this study, our dataset is changed in to Latin equivalent characters by using transliteration which is called GeezSERA 1.0 (published by Gasser[40]).

GeezSERA 1.0 is a system for ASCII representation of Ethiopic characters. This system support for romanizing Geez and geezifing romanized forms of African Semitic (Ethiopian Semitic) languages. Currently supported languages: Amharic, Tigrinya, Silte.

By default, romanization makes use of a modified version of the SERA conventions for romanizing Geez. Differences from standard SERA are as follows:

I. Words beginning with አ have '... instead of I... This is consistent with the representation of other characters and is a better reflection of the phonetics of Tigrinya.

II. Words beginning with other characters in the አ series begin with ' as well: ኦርቶዶክስ -> 'ortodoks

III. The vowel in the characters አ ዐ ሀ ሐ ኀ is "a" rather than "e", and the default form for geezifying combination like "'a", "ha", etc. is the first (Geez) form: አይደለም -> 'aydelem ሀገር, ሃገር -> hager hager -> ሀገር

IV. ` is replaced with ^ in `s, `h, and `S. This prevents ambiguity when the consonsant ` is followed by s, h, or S, as it can be in Tigrinya: ...ዕስ..., etc.

While there is no single agreed-on standard for converting Ge'ez script to Latin text, the GeezSERA 1.0 (Michael Gasser), which represents Ge'ez characters using ASCII characters is common in computational work on Ge'ez script and is used in this paper.

VI. Construction of general purpose stopword list

Like any other any other language Amharic writing system also contain different stopwords include prepositions, conjunctions, and articles. Even though these words are important in writing a document, they haven't advantage in designing NLP system.

Using these words in a dataset as it is have an impact on the performance the system, which mean, they would degrade its running speed and take much memory space while running. Due to this, removing stopwords from a dataset is necessary so as to reduce file size and processing time.

In this research, inorder to generate most common frequent stopwords from vacany announcment, a Python program was written to generate a Amhaic stopword list consisting of pronouns, prepositions, conjunctions and articles.

While in establishing a general stopword list, first we generate all the word forms appearing in a dataset are sorted according to their frequency of occurrence and the top most frequently occurring words are extracted, and then this list was examined manually to identify important word, for example the word “ልምድ” ranked at the 2nd position on the list and “የስራ” ranked at the 3rd position on the list but, these terms have their own role in extracting candidates from AVATs.

Finally, some non information-bearing words were included manually by the researcher even if they did not appear in the first top most frequent words. For example, various personal or possessive pronouns such as "ድርጅታችን", prepositions "በኋላ" and conjunctions "በተጨማሪ" were added.

VII. Sentence splitting

The Sentence Splitter module exports a simple sentence splitter for English. Given a string, assumed to be English text, it returns a list of strings, where each element is an English sentence. By default, it treats all occurrences of '.', '?' and '!' as sentence delimiters, but does its best to determine when an occurrence of '.' does not have this role (e.g. in abbreviations, URLs, numbers, etc.). Although this splitter is effective for English writing system, it is immposible to adapt straightforward to Amharic languages due to the punctuation marks difference from English writing system.

As aresult of this, the researcher developed a new algorithem for the puropse this reasearch and the "።" (አራት ነጥብ) used as sentence delimiters. It is important to emphasize that, this is a simple program is written only by considering only one punctuation mark,which is "።" (አራት ነጥብ) it didn't take in to consideration the other punctuation marks used as asentance delimiter in other document like "?" and "!", because, they are not recognized in AVAT as sentence delimiter.

```
Read raw of corpus
Char last = "።"
While (corpus !=null)
While (string != last)
    Concatenate string
Break the string
```

Fig. 4.5 Sentences splitter Algorithm

LEARNING AND EXTRACTION COMPONENT

This component is the fundamental part of our proposed model, which mainly deals with candidate texts. It uses the output of document preprocessing component as an input. The extraction and learning component also comprises different subcomponents that are used to make the data ready for extraction.

I. Part of speech tagger:

At this stage the tagger is going to assign a POS tag to each token. POS taggers have been applied to assign a single best POS to every word in a dataset. There are different part of speech tags of set in Amharic writing system, but for the purpose of this study we used 12 such as noun, noun phrase, verb, verb phrase, adjective, adverb, prepositions, punctuation, numeric, conjunction, verbal noun and noun consonant.

Since there is no POS tagged corpus available for AVAT the dataset was selected; preprocessed and 20 % from total words are manually tagged to train the POS tagger.

II. Candidate text selection:

Once the POS tagging process completed the next thing is identifying the possible candidate texts that would be extracted from the AVAT. The name of the organization, job position, Qualification, and agreement and the numbers which can represents salary, experience, number of people needed, phone and deadlines are considered as the candidate text, since the main motive of this research work is extracting such information from AVAT.

In order to identify and select the name of organization and job position in a AVAT text, Gazetteer is incorporated with the prototype system. The Gazetteer, which comprises list of different organization and job position names under consideration. The other candidates such as Qualification Salary, Agreement, Year of Experience, and Number of people needed, Deadline, and Phone are selected by analyzing their feature words.

New algorithm is developed, which used to extract the features of each candidate. The followings features are going to found: The current candidate word, previous/following of candidate word, the word before/after the previous /following word, POS of the above listed words, and the token category of the candidate token.

Finally after the candidate texts are identified from the dataset, they are tagged accordingly to their attributes. Here are the tags; those are used to tag the candidates:

<ORG> for organization name
<POS> for job position name
<QUL> for expected qualification in that position
<EXPER> for year of experience
<SAL> for salary

<AGREEMENT> for job Agreement

<NEED> for number of people needed

<DEAD> for deadline

<PHONE> for telephone

Read raw of corpus

Read gazetteer which contain list of organization name

Read gazetteer which contain list of position name

vacancy = each vacancy in raw of corpus

sting = tokens in vacancy

org = each organization name in gazetteer

pos = each position name in gazetteer

For vacancy

For string

If org == string

Tag the organization name by <ORG> at the beginning and </ORG> at the end of the organization name

End if

End for

For string

If pos == string

Tag the position name by <POS> at the beginning and </POS> at the end of the position name

End if

End for

If string == "ብዛት" and string + 1 == "<ADJ>"

Tag at end of the next word by <NEED>

If string == “የቅጥር” and string +1 == “<NP>” and string + 2 == “ሁኔታ”

If string == “የቅጥር” and string +1 == “<NP>” and “ኣይነት”

Tag at end of the next word the by <AGREE>

If string ==“ደመወዝ” and string+1 == “<VN>” or string == “ደሞዝ” and string+1 == “<VN>

Tag at end of the next word the by <SAL>

If string == “ኣመት” and string +1 == “<NUMP>” and string +2 == “ከዚያ” and “<PRONP>” and “በላይ”

Tag the word befor “ኣመት” by <EXPER>

Tag the word after “ኣመት” by </EXPER>

Elseif string ==“ኣመት” and string +1 == “<NUMP>” and string +2 == “የስራ”and string +3 == “<NP>” and “ልምዲ”

Tag the word befor ”ኣመት” by <EXPER>

Tag the word after ”ኣመት” by </EXPER>

Elseif string ==“ኣመት” and string +1 == “<NUMP>” and string +2 == “የስራ”and string +3 == “<NP>”

Tag the word befor ”ኣመት” by <EXPER>

Tag the word after ”ኣመት” by </EXPER>

End if

If string == “ተከታታይ” and string +1 “<ADJ>” and “የስራ”and string +2 == “<NP>” and “ቀናት”

Tag the word befor ” ተከታታይ” by <DEAD>

Tag the word after ” ቀናት” by </DEAD>

Elseif string == “ተከታታይ” and string +1 “<ADJ>” and string+2== “ቀናት”

Tag the word befor ” ተከታታይ” by <DEAD>

Tag the word after ” ቀናት” by </DEAD>

Elseif string “የሰራ”and string +1 == “<NP>” and “ቀናት”

Tag the word befor ” ተከታታይ” by <DEAD>

Tag the word after ” ቀናት” by </DEAD>

End if

If string ==“ስልክ” and string+1 == “<N>” and “ቁጥር”

Tag at end of the next word the by <PHONE>

Elseif string == string ==“ስልክ” and string+1 == “<N>” or “መረጃ”

Tag at end of the next word the by <PHONE>

Elseif string == “ለበለጠ” and string +1 ==“መረጃ” and string+2 == “<N>” and string +4 ==<NUMCR>

Tag at end of the next word the by <PHONE>

End if string == “የትምህርት”and string + 1==<NP> or “ተፈላጊ” and string + 1 == <ADJ> and

If string == “ደረጃ”and string +1== <ADJ> or “ችሎታ” and string +1== <ADJ>

Tag the word after “ደረጃ” by <QUL> or

Tag the word after “ችሎታ” by <QUL>

Elseif string == “የተመረቀች” and “<PUNC>”

Tag at end word “የተመረቀች” by </QUL>

Elseif string == “ዲግሪ” and “<PUNC>”

Elseif string == “ዲፕሎማ” and “<PUNC>”

Tag at end word “ዲፕሎማ” by </QUL>

Elseif string == “ሰርተፊኬት” and “<PUNC>”

Tag at end word “ሰርተፊኬት” by </QUL>

End if

End for

Fig. 4.6 Candidate text identifier and tagger Algorithm

III. Candidate text extraction:

Once the intended candidates are identified and tagged in candidate text selection phase, extraction of those candidate texts are done consecutively to their category. The other data those not selected by the system as a candidate text from AVAT would be discarded. A rule-based algorithm is developed, which aided to extract such a tagged candidate texts from a dataset. Here is an algorithm:

```
Read raw of corpus
Vacancy = each Vacancy in raw of corpus
String = each tokens in vacancy
For string
    If string is tagged by <ORG>
        Hold the position
        While( string != </ORG>)
                Print the string
                Increament string
        End while
    End if
    If string is tagged by <POS>
        Hold the position
        While( string != </POS>)
                Print the string
                Increament string
        End while
    End if
```

```
    If string is tagged by <EXPER>
        Hold the position
        While( string != </EXPER>)
                Print the string
                Increament string
        End while
    End if
      If string is tagged by <QUL>
        Hold the position
        While( string != </QUL>)
                Print the string
                Increament string
        End while
    End if
If string is tagged by <DEAD>
        Hold the position
        While( string != </DEAD>)
                Print the string
                Increament string
        End while
    End if
If string is tagged by <SAL>
        Print the token
Elseif  string is tagged by <NEED>
        Print the token
Elseif  string is tagged by <AGREE>
        Print the token
Elseif  string is tagged by <PHONE>
        Print the token
End if
```

Fig. 4.7 Candidate text extractor Algorithm

POST PROCESSING

This is the last component of the proposed model. After the relevant information has been founded by applying the extractor algorithm on the given dataset the extracted candidate text fragments are assigned to the corresponding attributes of the target structure and store them in the database according to the predefined format of the database slots. In this research work the nine attributes that are extracted are stored in the table as follows. Thus the main function of the post processing component is to arrange the format and store the extracted data in a database, so that, it will be flexible for data mining or any other application which want to use the data.

The extracted candidate texts are also normalized according to the expected format, since, some identified facts may appear in text more than once and there might be violation the properties of the database.

THE PROTOTYPE SYSTEM

As we have supposed in above, once the candidate texts are extracted and stored in a database it is easily applicable for different applications. For the purpose of this research the prototype system which called AVATIES (Amharic Vacancy Announcement Text Information Extraction System) is developed to facilitate a user search in regarding Amharic language vacancy announcement. For the development of the prototype AVATIES Microsoft visual basic programming and Microsoft SQL server 2008 are employed. Visual basic programming was also employed while in developing algorithm which used for accessing data from database. The following figure 4.8 shows AVATIES.

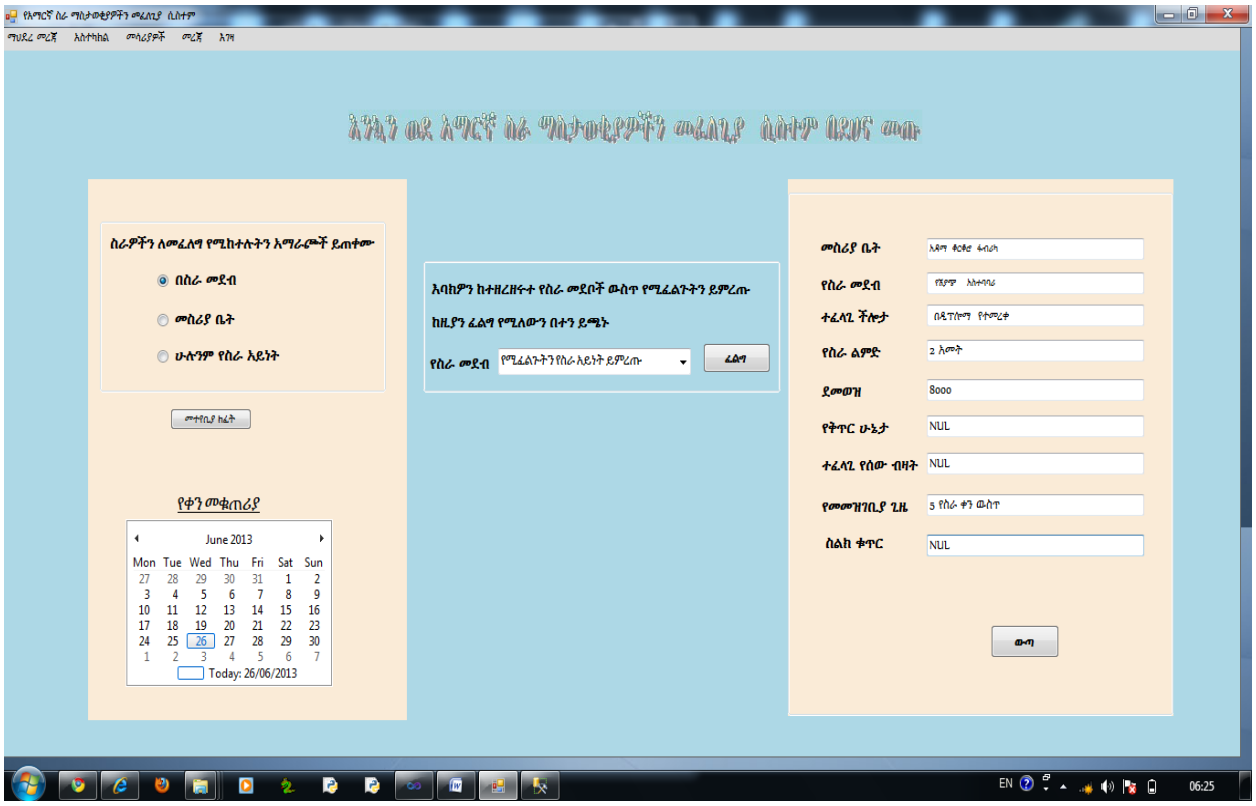

Figure 4.8 User interface after the data is extracted from the database

CHAPTER FIVE
RESULT AND EVALUATION

5.1. INTRODUCTION

Natural language systems are designed and developed to perform specific tasks as required and expected by users or other systems. A machine translation system is expected to give a correct translation for a given input. An information retrieval system (search engine) is expected to retrieve correctly ranked relevant documents and part of speech tagger is expected to assign a correct tag to a given instance of a word. Similarly, Information Extraction system is also expected to extract the right information from a text.

In general, for a given input, the NLP system is expected to give a correct output. What constitutes correct output and how we can measure it is, however, not an easy task and so is an active area of research in natural language processing. For example, given that two human translators do not translate the same Amharic text into the same English text, how can a translation produced by a machine be measured for correctness? The answer is not trivial. Similarly, how can we measure the correctness of information extraction system, which is naturally an easier problem than machine translation? Apart from output correctness, there are other issues to raise about NLP systems: how easy are they to use by non-experts, how well do they plug into other components, can they be ported and maintained by someone who did not participate in the system development? Therefore, raising one or more questions of accuracy, user-friendliness, efficiency, modularity, portability and robustness is important depending on the purpose.

Evaluation is the process of measuring one or more of the above qualities of an algorithm or a system. It has an important role in natural language processing for both system developers and technology users.

With evaluation, system developers are much better equipped with the knowledge of what components to improve in the system in order to achieve the desired goals. It helps researchers communicate their results and compare them with previous research work. For users, evaluation provides them with the necessary information they need to easily compare alternative systems and choose the one that meets their requirements.

In this section, we will discuss the kind of evaluation methods used in reporting and analyzing information extraction results.

5.2. EVALUATION METRICS

In this thesis, we will do mainly an intrinsic, black-box and automatic evaluation. We evaluate the different information extraction algorithms as isolated systems (intrinsic). Within the isolated system, we are going to do black-box evaluation as we will only compare the outputs of the system for given inputs with the gold standard.

The most commonly used evaluation metrics in information extraction are accuracy, error rate, precision, recall and f-measure.

Accuracy is the ratio of the number of correct outputs to the total number of outputs for given inputs. Error rate is the ratio of errors to the total number of outputs.

Mathematically,

Accuracy = number of correct outputs/ number of total input-output pairs

Error rate = number of incorrect outputs/number of total input-output pairs

Precision and recall are concepts first widely used in evaluation of information retrieval systems.

In IR, precision is defined as the number of relevant documents retrieved by a search engine divided by the total number of documents retrieved by that search, and recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents including those which should have been retrieved.

Precision and recall can be used in evaluating information extraction as well. In this case, precision is the number of items correctly labeled as belonging to the class of interest (true positives) divided by the total number of items labeled correctly or incorrectly as belonging to that class (true positives + false positives). Recall, by contrast, is the number of items correctly labeled items (true positives) divided by the total number of items that actually belong to the class which includes items not correctly identified as belonging to that class (false negatives).

Mathematically,

$$P = \frac{TP}{TP+FP} \quad \text{.....(6.1)}$$

$$R = \frac{TP}{TP+FN} \quad \text{.....(6.2)}$$

Precision and recall are inversely related. When one is increased, the other decreases. For example, in information extraction, if we are focusing on organization slot and all organization are extracted correctly, and then we get a 100% recall, whereas the precision decreases significantly as there are many organization are not extracted correctly.

Usually, precision and recall are combined to give one value called the f measure. The f-measure is the weighted harmonic mean of precision and recall.

Mathematically,

$$F = \frac{2PR}{P+R} \quad \text{.....(6.3)}$$

Equation 6.3 is the harmonic mean of precision and recall, where both measures are given equal importance.

5.3. THE DATASETS

It is mentioned that in section 1.4.2, in order to design a robust IE system, a variety of training data is required to ensure existence of intended candidate entities in each vacancies.

The dataset used for this work was AVAT acquired from the "Ethiopian reporter" newspaper published in Amharic twice in week. For the purpose of this study, 116 AVATs that contain in general 10,766 words were selected purposely with different range of vacancy announcements. There dissimilarity is based on the organization of who is posted the vacancies and the type of vacancies.

	Training data	Test data
Number of vacancies	82	34
Number of word (tokens)	8,046	2,720
Number of organization data	82	34
Number of job position data	82	34
Number of qualification	82	34
Number of salary data	78	32
Number of people needed data	62	26
Number of experience data	76	29
Number of deadline data	82	34
Number of phone data	69	28

Table 5.1 the Statistics for dataset used

5.4. EXPERIMENTAL RESULT AND EVALUATION EACH COMPONENT OF OUR SYSTEM

5.4.1. EXPERIMENTAL RESULT AND EVALUATION OF NORMALIZATION

The performance of our system has been evaluated before and after document normalization. The experimental result showed that document normalization has a significant effect on the performance of the system. Consider figures 5.1 and 5.2 to see the impact of document normalization:

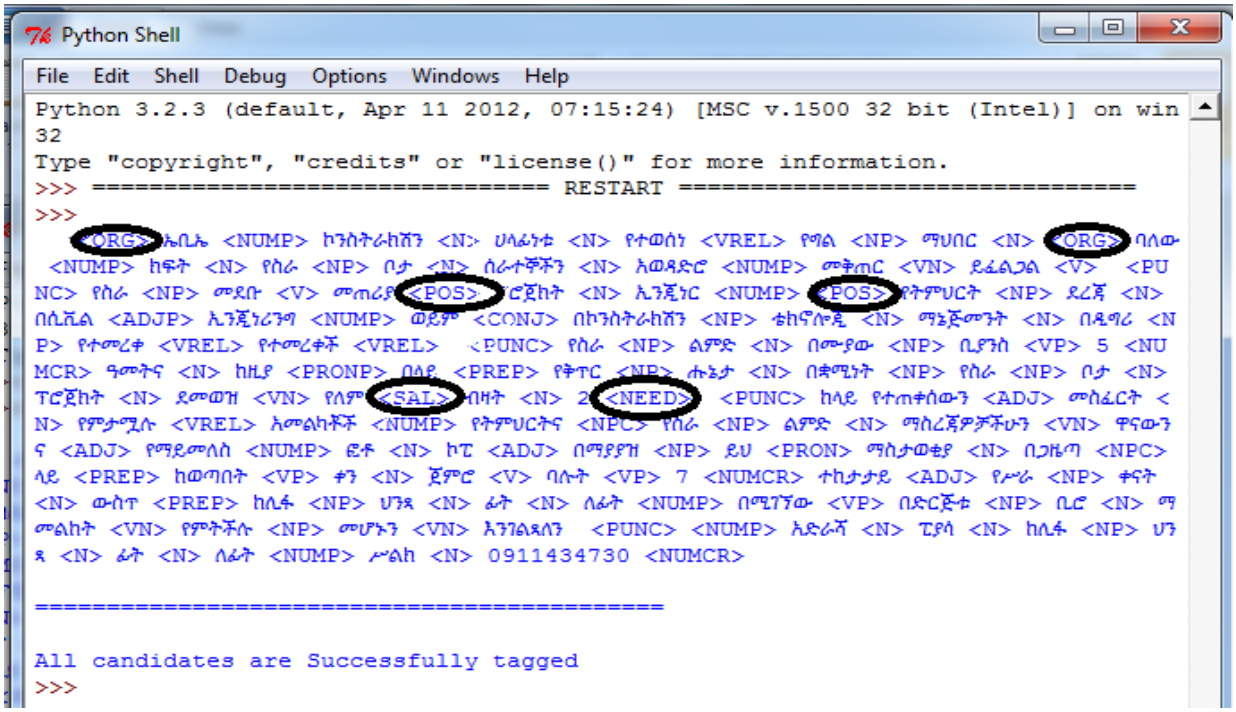

Figure 5.1 Before Normalization

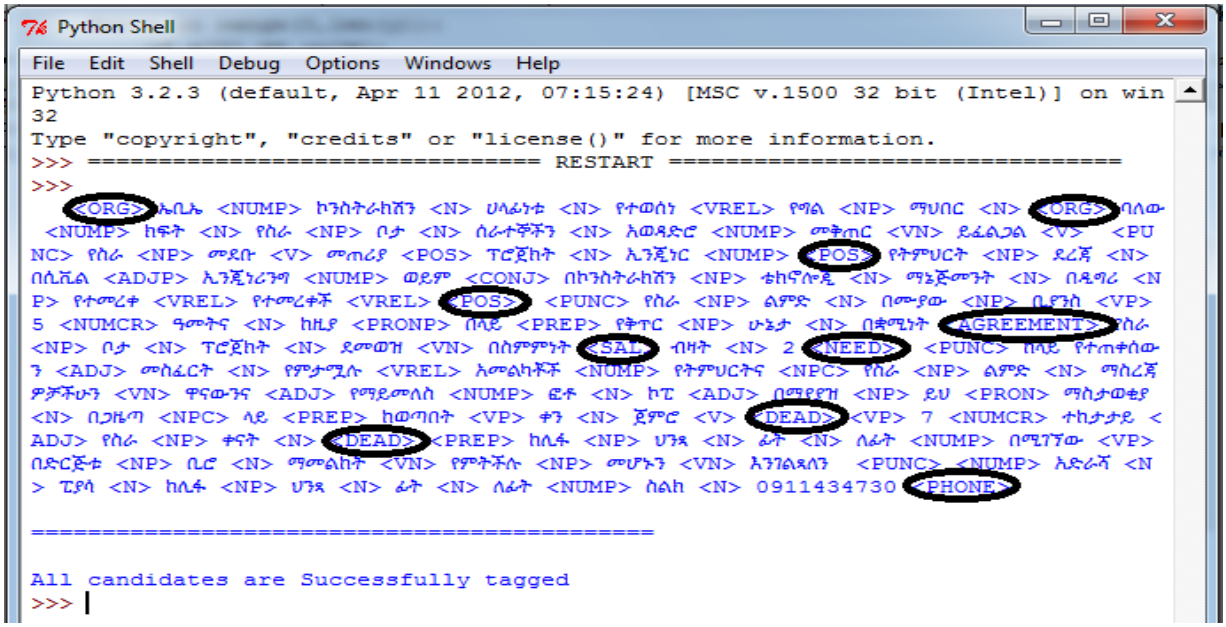

Figure 5.2 After Normalization.

The above two figures; Figure 5.1 and 5.2 showed a sample result of candidate texts for AVAT. As the first result illustrated that, for a single AVAT, the system considered only four data as a candidate text for that vacancy announcement, but in the second result, the candidate texts are augmented in to seven for the same vacancy announcement, this is due to normalization. Thus, before provided any type AVAT to AVATIES, it should be normalized.

5.4.2. EXPERIMENTAL RESULT AND EVALUATION OF STOPWORD REMOVAL

As discussed in section 4.3, using stopwords in dataset as it is has an impact on the performance of the system.

Datasets	Before stopwords are removed	After stopwords are removed
	Time	Time
	2:42 min	1:36 min

Table 5.2. Effects of stopword removing on the running speed of the system

As Table 5.2 shows that running speed of the system is increased by 1.06 minute than before stopwords were removed. Still the running time indicated that it could improve, if all unnecessary words are removed. Nevertheless, in this research, it is impossible to say all stopwords were included, when stopword list is constructed.

5.4.3. EXPERIMENTAL RESULT AND EVALUATION OF TRANSLITERATION

It is discussed in section 4.3, the main reason why we used transliterated dataset is POS tagger that we used in this study is works only for Latin characters. Once the dataset is tagged, it should be re-transliterate or back to gee'z type, before it is going to another processing step. So in order to do this, another algorithm is required. For the purpose this research, two types of transliterater was tasted.

Latin document	Correctly re-transliterate words	Incorrectly re-transliterate words
Number of word	205	1958

Table 5.3 experimental result of transliterater that developed by the researcher.

Latin document	Correctly re-transliterate words	Incorrectly re-transliterate words
Number of word	2138	67

Table 5.4 experimental result of transliterater that developed by Gaser

The experimental result shows that the second one could score 97.3% accuracy and the second transliterater shows 9.3% accuracy, due to this, the second one is chosen for this research.

5.4.4. EXPERIMENTAL RESULT AND EVALUATION OF PART OF SPEECH TAGGER

Now a days, there are different types of NLP tools are available, though, not all tools are fully relevant for Amharic language. Among this POS tagger is a one tool which is commonly used in designing most of natural language proceeding system [36].

For the purpose of this study two statistical POS taggers were tested, the first one is Brill POS tagger for Amharic language, which was developed by Gebrekidan [36]. Bigram POS tagger is another tagger that we have tasted in this study, which is developed by Abebe [37].

	Correctly tagged	Incorrectly tagged
Number of words	1565	640

Table 5.5 Experimental result of Bigram POS tagger

	Correctly tagged	Incorrectly tagged
Number of words	1962	243

Table 5.6 Experimental result of Brill POS tagger

From the above table what we can understand is Brill POS tagger has 89.5% of accuracy and Bigram POS tagger has an accuracy of 71.4 %. Hence, the researcher selected and used Brill POS tagger for tagging the dataset.

The researcher had tried to tag the correct part of speech tags for each word manually for those that are incorrectly tagged while in part of speech tagging process using Brill POS tagger.

5.4.5. EXPERIMENTAL RESULT AND EVALUATION OF PROTOTYPE SYSTEM FOR CANDIDATE TEXT EXTRACTION

Evaluation of candidate text extraction may seem relatively straightforward: first preprocessed a datasets then tagged with POS; the candidate text extraction is run on this dataset; we count correct, missing, and spurious candidates (as we did for part of speach tagging) and then compute recall, precision, and F measure. Understandably, the scores will depends on the dataset and the set of candidate text types.

5.4.5.1. EXPERIMENTAL RESULT AND EVALUATION OF ORGANIZATION AND POSITION EXTRACTION

It is obvious that, placing organization name and position name is obligatory for any type of vacancy announcement. Extracting these candidate texts with other candidate texts from AVAT was the main objective of this study.

Two algorithms have been tasted to handle and extract organization and position candidate texts from AVAT. The first algorithms is based on feature words or context information, which means extracting candidates based on the neighborhood features words those can express the name organization and position. Gazetteer based identification and extraction was another algorithms that the researcher had tested. We evaluate the performance of our system for identification and extraction by using two known evaluation mechanisms in NLP, they are Recall and precision. In this case, Recall is the proportion of candidate texts which are extracted correctly over the total number of extracted candidates for each slot in the test dataset. Likely, precision is the proportion of candidate texts which are identified and extracted correctly over the number of identified and extracted for each slot in the test dataset.

	Recall	Precision	F-measure
Organization	47.8	64.4	54.8
Position	36.9	56.2	43.19

Table 5.7 Experimental result of context information based Algorithm for Organization and Position extraction

	Recall	Precision	F-measure
Organization	100	100	100
Position	100	100	100

Table 5.8 Experimental result of gazetteer based Algorithm for Organization and Position extraction

The experimental result showed that integrating gazetteer with organization and position extractor algorithm could have an ability to improve the performance of the system.

The main reason why feature based organization and position identification and extraction algorithm was not good as gazetteer based algorithm is that: in different AVAT both organization and position presented in several ways, which means their presentation likeness from one AVAT to another is very rare. Therefore, it is not possible to handle and discover all those various ways of representation based on feature words or context information. As a result of this, the second algorithm was not that much effective in identifying and extracting organization and position name as gazetteer based algorithm.

5.4.5.2. EXPERIMENTAL RESULT AND EVALUATION OF OTHER CANDIDATE TEXT EXTRACTION

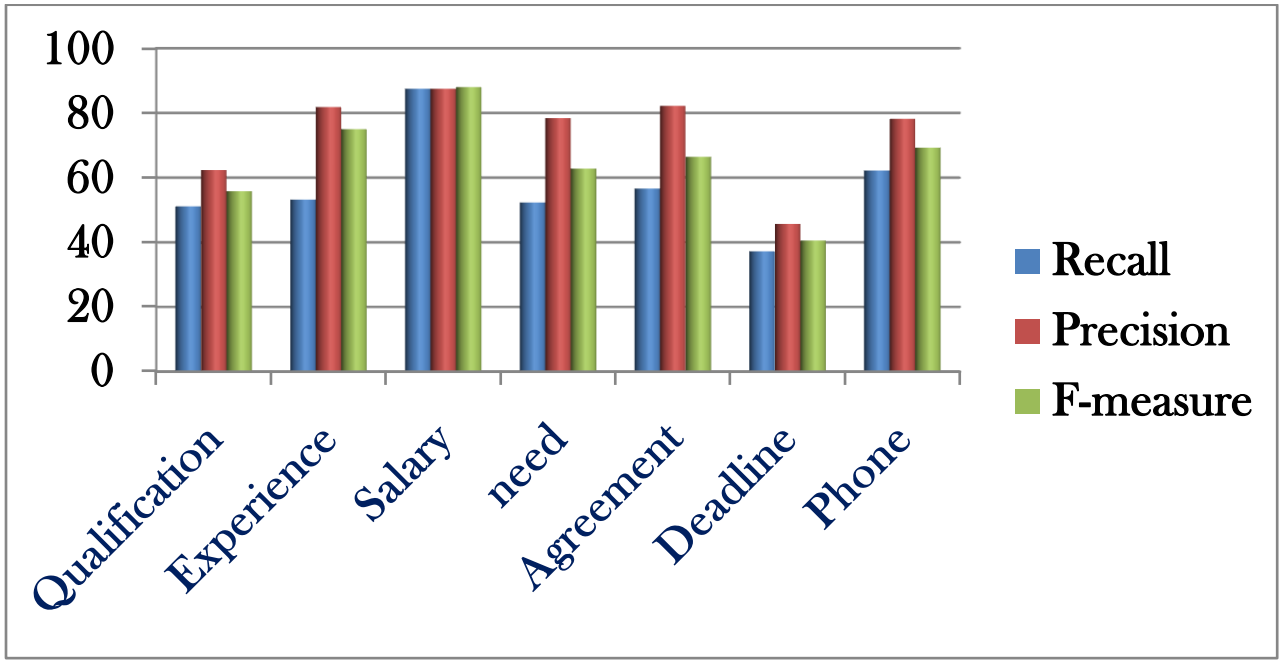

Fig 5.9 Experiment result of the rest candidate text extraction

The candidates such as “Salary”, “Number of needed people”, “Agreement” and “Phone” are provided the best performance. It might be due to the facts those used for representing these candidate texts in AVAT usually uses the same pattern: For example, job agreement is presented in most of the time in the fowllowing format: <የቅጥር ሁኔታ> or <የቅጥር አይነት> expected word are “በቋሚነት”, “ቋሚ”, “በኮንትራት”, “ኮንትራት”.

The worst performer was “Quliifcation” and “Deadline” slot. The main reason was over generalization in specific selcetion rule: "የትምህርት ደረጃ" * "የተመረቀ ወይም የተመረቀች and <PUNC>" or “ተፈላጊ ችሎታ” * “ዲግሪ and <PUNC>". This rule is meant to match “የትምህርት ደረጃ ከዩኒቨርሲቲ በግብርና ሰፕላይስ ማኔጅመንት በኢኮኖሚክስ በቢኤ ዲግሪ የተመረቀ ወይም የተመረቀች <PUNC” and “ተፈላጊ ችሎታ በአካውንቲንግ የመጀመሪያ ዲግሪ <PUNC>” rspectively but it also matches the wrong sentence like “በአካውንቲንግ የመጀመሪያ ዲግሪ ያለው” or “ዲግሪ ያለው”. We need to inspect more vacancy announcement documents in order to refine selection rule and to improve our system performance.

Generally, the result of experiment shows that, our system can still be improved. Although this algorithm shows good result on precision, that is 79.56%, a cumulative recall is lower at 66.6% and F-measure was 71.7%. Low recall is common in most of IE. Using job domain document, RAPIER Calif [10] had precision 84% and recall 53%.

CHAPTR SIX
CONCLUSION AND RECOMMENDATION

6.1. CONCLUSIONS

In this study we presented the first rule-based IE system for AVAT. The following conclusions are drawn from the experiments with regard to the research questions:

- The results obtained from experiments shows 79.56% precision and 66.6% recall on 34 AVAT test dataset.
- Contextual features (previous, next word, two word previous, and two words next to current word) along with POS of each word are very useful for AVATIES to predict candidates in a text.
- Absence of contextual word for some candidates in a dataset is the major problem, since the algorithm extracts candidate texts based on this feature.
- The AVATIES does not give similar accuracy on different datasets. The accuracy depends on the features found in each candidate. Some candidates may contain similar feature words, that incorporated within the rule and other may contains feature words out of these, therefore their accuracy may differ depending on the test dataset.
- The experiments have been carried out on each component of a system separately, in order to evaluate them individually. Based on the result, candidate text selector algorithm has shown less accuracy as compared with other components, this due to lack of adequate rules or feature words for each candidate text.
- Extracting candidates is a challenging task in rule-based algorithm, because one candidate text may appear in various ways in different AVAT. In this research various rule was tried to incorporate in candidate text selector Algorithm, which used to identify candidate text in AVAT. So, we can confidently say that it is promising to develop an IE system using the knowledge engineering approach.

6.2. RECOMMENDATION

The IE system proposed in this research work is not complete it requires different improvements to make the system more robust and to be used at a large scale. The following recommendations are forwarded for future work:

- The dataset used in this system is only from one newspaper. Using a sizable dataset from different newspaper could possibly help to get diversified rules and improved performance.
- Further research is expected on different IE tools, such as sentence Parser, POS, NER, and Co-reference Resolution for Amharic language to develop an effective IE system.
- It would be interesting to implement a statistical algorithm to identify and extract candidate texts and test to see how it performs for AVAT.
- According to Ekbal and Bandyopadhyay [39] in order to design effective IE system, NER is an irreplaceable tool. A gazetteer, which contains different names collected from the AVATs and other sources, is used. For recognizing names, using automatic named entity recognition in later stages could minimize the burden of selecting the named entities.
- In order to make the system practicable, a different rule should be incorporated in to the algorithms and further improvements on the system are required to augment the running speed of the system.

REFERENCE

[1]. Philipp Johannes, Multilingual Information Extraction, Department of Computer Science, University of Helsinki 15th, February 2004.

[2]. Ellen Riloff, Inducing Information Extraction Systems for New Languages via Cross-Language Projection, School of Computing, University of Utah, Salt Lake City, June 2004.

[3]. Jim Cowie and Yorick Wilks, Information extraction, Lecture note on Information extraction, 2007.

[4]. Cunningham H., Information Extraction Automatic, Encyclopedia of Language & Linguistics journal, Second Edition, volume 5, pp.665-677, Oxford, Elsevier, 2006.

[5]. Shubin Zhao, Information Extraction from Multiple Syntactic Sources, A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Department of Computer Science, New York University, May 2004.

[6]. Christian Siefkes and Peter Siniakov, an Overview and Classification of Adaptive Approaches to Information Extraction, LNCS Journal on Data Semantics IV, 2005.

[7]. Teklu Surafel, Automatic classification of Amharic News text, A thesis submitted in partial fulfillment of the requirements for the degree of Masters of science, Department of information Science Addis Ababa University, June 2003.

[8]. Y. Wilks and C. Brewster, Natural language processing as a foundation of the semantic web, Foundations and Trends in Web Science, 2000.

[9]. Katharina Kaiser and Silvia Miksch, Information Extraction, A survey, Institute of Software Technology & Interactive Systems, Vienna University of Technology, May 2005.

[10]. Mary Elaine Califf and Raymond J. Mooney, relational learning pattern-matching rules for Information Extraction, Department of Computer Sciences, University of Texas at Austin, July 1987.

[11]. Douglas E. Appelt and David J. Israel, Introduction to Information Extraction Technology, A Tutorial Prepared for IJCAI-88, Artificial Intelligence Center SRI International 333 Ravenswood Ave, Menlo Park, CA.

[12]. Hovy, E. Automated Text Summarization, The Oxford Handbook of Computational Linguistics, Oxford University Press, pp.583-598, 2005.

[13]. Seid Muhie Imam, Amharic Question Answering (AQA), Department of Information Technologies, Adam University, Adam, 2002. Available at http:// www.acadamia.com

[14]. Dipanjan Das and Andre F.T. Martins, A survey on automatic text summarization, Language Technologies Institute Carnegie Mellon University, Nov 2007.

[15]. Line Eikvil, Information extraction from World Wide Web, a survey, July 1999.

[16]. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, An Introduction to Information Retrieval, Cambridge University, England, April 2009.

[17]. Getasew Tsedalu, Information Extraction for the Amharic News Text, A thesis submitted in partial fulfillment of the requirements for the degree of Masters of science, Department of Computer Science Addis Ababa University, November 2010.

[18]. Hamish Cunningham, Information Extraction Automatic, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK.

[19]. Fabrizio Sebastiani, Text Categorization, Dipartimento di Matematica Pura, Applicata Universit, AdiPadova 35131 Padova, Italy, 2002.

[20]. Dwi H. Widyantoro and Yudi Wibisono, Information Extraction for E-Job, Marketplace,School of Informatics and Electrical Engineering ITB.

[21]. Alberto Tellez-Valero, Manuel Montes-y-Gomez and Luis Villasenor-Pineda, Using Machine Learning for Extracting Information from Natural Disaster News Reports, CIC Ling 2005.

[22]. Benjamin Rosendfeld, Ronen Feldman, Moshe Fresko, TEG—a hybrid approach to information extraction, Knowledge Information Systems, pp. 1–18, 2005.

[23]. Ferran Pal and Antonio Molina, Natural language engineering: improving Part of speech tagging using lexicalized HMMs, Cambridge university press, united kingdom, 2004.

[24]. Kameyama M., Information Extraction across Linguistic Barriers, AAAI Spring Symposium Series on Cross-Language Text and Speech Retrieval, Stanford, 1997.

[25]. Ying Yu, Xiao-Long Wang, Yi Guan, Information Extraction for Chinese Free Text Based On Pattern Match Combine with Heuristic Information, In Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, pp 4-5 November 2002. Available at http://news.cina.com.cn/special/gov/index.shtml

[26]. Matthew W. Bilotti, Boris Katz, and Jimmy Lin, What Works Better for Question Answering: Stemming or Morphological Query Expansion, Massachusetts Institute of Technology, Cambridge, USA, 2004.

[27]. Tomek Strzalkowski and Sanda Harabagiu, Advances in Open Domain Question Answering, Published by Springer, Netherlands, 2008.

[28]. Jochen L. Leidner and Chris Callison-Burch, Evaluating Question Answering Systems Using FAQ Answer Injection, In Proceedings of the 6th Annual CLUK Research Colloquium, 2003.

[29]. Juan Antonio P´erez-Ortiz and Mikel L. Forcada, Part-of-Speech Tagging with Recurrent Neural Networks,Departament de Llenguatges Sistemes Inform`atics Universitat d'Alacant E-03071 Alacant, Spain, 2001.

[30]. Samuel Eyassu and BjörnGambäck, Classifying Amharic News Text Using Self-organizing Maps, Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan, 2005.

[31]. Getachew Haile, the Problems of the Amharic Writing System, Unpublished, 1966.

[32]. M. Bender, J. Bowen, R. Cooper, and C. Ferguson, Languages in Ethiopia, Oxford Univ. Press, London, 1976.

[33]. ባየይማም፤ የአማርኛስዋሰው፤ት.መ.ማ.ድ፤1987.

[34]. Bizuneh Mamuye Birhan, the application of Websom for Amharic Text Retrieval, A Thesis submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment for the Degree of Master of Science in Information Science, July 2003.

[35]. Martha Yifiru Tachbelie and Wolfgang Menzel, Amharic Part-of-Speech Tagger for Factored Language Modeling, Department of Informatics, University of Hamburg ,Hamburg, Germany.

[36]. Binyam Gebrekidan, Natural Language Processing & Human Language Technology, Part of Speech Tagging for Amharic, A Project submitted as part of a program of study for the award of MA, School Of Law, Social Sciences and Communications, United Kingdom, June 2010.

[37]. Ermias Abebe, Bigram part-of-speech tagger, Addis Ababa University, school of Information Science, Addis Ababa, 2006.

[38]. Ralph Grishman, Silja Huttunen and Roman Yangarber, Information extraction for enhanced access to disease outbreak reports, Journal of Biomedical Informatics 35(4): 236-246, 2002.

[39]. Ekbal, A. & Bandyopadhyay, NER Using Support Vector Machine: A Language Independent Approach, International Journal of Electrical and Electronics Engineering, pp. 155-170, India-700032, 2010.

[40]. Michael Gasser, Horn Morpho: Morphological analyzer and generator for Amharic and Oromiffa language, Indiana University, USA, 2011.

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 material used for the thesis has been duly acknowledged.

Sintayehu Hirpassa

June 2013

This thesis has been submitted for examination with my approval as university advisor

---

Ermias Abebe (Mr)

June 2013

Appendix I: Sample test data

ክፍት የሥራ መደብ ማስታወቂያ በመከላከያ ኮንስትራክሽን ኢንተርፕራይዝ ከዚህ በታች በተዘረዘሩት ክፍት የስራ መደቦች አመልካቾችን አወዳድሮ መቅጠር ይፈልጋል። የሥራ መደቡ መጠሪያ ቢሮ አስተዳደር ኦፊሰር ተፈላጊ የትምህርትና የሥራ ልምድ ዲፕሎማ (10+3) በጽሕፈትና ቢሮ አስተዳደር እና ከ4-5 ዓመት የሥራ ልምድ ወይም 10+2 ቴክኒክና ቮኬሽናል ዲፕሎማ እና 6 ዓመት አግባብ ያለው የሥራ ልምድ ያላት ብዛት 5 የሥራ ቦታ አ/አ ዋናው መ/ቤት ደመወዝ በስምምነት የማመልከቻ ቦታ በምድር ኃይል የሰው ኃይል አስተዳደር ቢሮ ለበለጠ መረጃ፡- ስልክ ቁጥር 011-8962016 አመልካቾች ከላይ ለተጠቀሱት ክፍት የሥራ መደቦች ዋናውንና ኮፒውን የትምህርትና የስራ ልምድ መረጃቸሁን በመያዝ ማስታወቂያ ከወጣበት ዕለት አንስቶ ባሉት 10 ተከታታይ የስራ ቀናት ቀርባችሁ መወዳደር የምትችሉ መሆኑን እናሳውቃለን።

የኢትዮጵያ ምርት ገበያ ባለሥልጣን ክፍት የሥራ ቦታ ማስታወቂያ የሥራ መደቡ መጠሪያ የሪከርድና ማህደር ባለሙያ ዝቅተኛ ተፈላጊ ትምህርት፣ ክህሎትና የሥራ ልምድ ትምህርት በማኔጅመንት ወይም በማኔጅመንት ኢንፎርሜሽን ሲስተም የኮሌጅ ዲፕሎማ ቀጥታ አግባብ ያለው የስራ ልምድና ክህሎት 4 ዓመት ሥራ ልምድ ደረጃ VI ደመወዝ 2,154 የሥራ መደቡን ዝቅተኛውን ተፈላጊ ችሎታ የምታሟሉ ወይም ከዚያ በላይ ትምህርት እና የሥራ ልምድ ያላችሁ አመልካቾች ማስረጃዎቻችሁን ዋናዎቹን ከማይመለስ አንድ አንድ ፎቶ ኮፒ ጋር በመያዝ ሜክሲኮ አደባባይ በሚገኘው ዋናው መ/ቤት የሰው ሃብት ሥራ አመራርና ልማት ዳይሬክቶሬት ሂደት 3ኛ ፎቅ ቢሮ ቁጥር 302 መመዝገብ የምትችሉ መሆኑን እንገልፃለን። የምዝገባ ጊዜ የካቲት 13 እስከ 19 ቀን 2005 ዓ.ም ድረስ ዘወትር በሥራ ሰዓት። ለተጨማሪ መረጃ በስልክ ቁጥር 0115-158181 ወይም 0115-537122 ወይም 0115-512734 ደውሎ መጠየቅ ይቻላል።

ክፍት የሥራ ቦታ ማስታወቂያ ሴንቸሪ ጄኔራል ትሬዲንግ ኃ.የተ.የግ.ማኅበር /ሴንቸሪ ፕሮሞሽን ሰርቪስ / የንግድ ትርኢቶችን ኤግዚብሽኖችን ወርክ ሾፖችን እና ሌሎች ተዛማጅ ፕሮግራሞችን በማዘጋጀት ይታወቃል። ድርጅቱ ከዚህ በታች በተመለከቱት ክፍት የሥራ ቦታዎች አመልካቾችን አወዳድሮ በቋሚነት ለመቅጠር ይፈልጋል።የሥራ የሥራ መደቡ መጠሪያ ሹፌር ተፈላጊ ችሎታ ከስምንተኛ ክፍል በላይ 3ኛ እና ከዚያ በላይ መንጃ ፍቃድ ያለው ከ 3 አመት በላይ የስራ ልምድ ያለው ደመወዝ በስምምነትብዛት 1 ከላይ የተጠቀሰውን መመዘኛ የምታሟሉ አመልካቾች የትምህርት፣የሙያ እና የስራ ልምድ ማስረጃችሁን የማይመለስና ፎቶ ኮፒ ከማመልከቻቸሁ ጋር በማያያዝ ይህ ማስታወቂያ ከወጣበት ቀን ጀምሮ ባሉት 10/አስር/ የሥራ ቀናት ውስጥ ወሎ ሰፈር ሚና ህንፃ 2ኛ ፎቅ ላይ በሚገኘው የድርጅቱ ዋና መ/ቤት በሥራ ሰአት በመገኘት መመዝገብ የምትችሉ መሆኑን እናስታውቃለን ስልክ፡ 011-5-549245/46 0910-016330

ክፍት የሥራ ቦታ ማስታወቂያ ሴቡ ትሬዲንግ ኃ/የተ/ የግል ማኅበር ድርታችን ሴቡ ትሬዲንግ ኃ/የተ/ግል/ ማኅበር ከባድ የኮንስትራክሽን መሣሪያዎችን በማከራየትና በአልሙኒየም ሥራ ላይ የተሰማራ የግል ድርጅት ሲሆን ከዚህ በታች ለተመከተው ክፍት የሥራ ቦታ ሠራተኞችን አወዳድሮ ለመቅጠር ይፈልጋል።የሥራ መደቡ መጠሪያ ሹፌር ደመወዝ በስምምነት ብዛት 2 ተፈላጊ ችሎታ 10ኛ ክፍል ያጠናቀቀ እና 3ኛ ደረጃ መንጃ ፈቃድ ያለው እንዲሁም ከአ.አ ውጭ ከፍለገር የመንዳት ልምድ ያለው የሥራ ልምድ ከ3 ዓመት በላይ። የመመዝገቢያ ጊዜ፡ ከላይ የተገለጸውን መስፈርት የምታሟሉ አመልካቾች የትምህርት ማስረጃችሁን እና የሥራ ልምድ ዋናውንና የማይመለስ ኮፒ ከአንድ ጉርድ ፎቶ ግራፍ ጋር በመያዝ የድርጅቱ ቢሮ በሚገኝበት መካኒሳ አቦ ቤተክርስቲያን አካባቢ ባለው አስመላሽ ሕንጻ 2ተኛ ፎቅ ቢሮ ቁ. 792/በ ድረስ በመምጣት መመዝገብ የምትችሉ መሆኑን እንገልጻለን።ለበለጠ መረጃ በስ.ቁ. 251 113 200143/+251 913 294164

ክፍት የሥራ ቦታ ማስታወቂያ ቢቲ ዲጅታል ቢዝነስ ብሪጅ ኃላ.የተ.የግ.ማኅበር ከዚህ በታች የተዘረዘረውን መስፈርት የሚያሟሉ ሥራ ፈላጊዎችን አወዳድሮ ለመቅጠር ይፈልጋል። ስለዚህ አመልካቾች ይህ ማስታወቂያ ከወጣበት ቀን ጀምሮ ባሉት 10 ተከታተይ የሥራ ውስጥ ዋናውንና የማይመለስ ፎቶኮፒ የትምህርትና የሥራ ልምድ ማስረጃዎችን በመያዝ ላፍቶ መብራት ኃይል አደባባይ ወደ ሃና ማርያም አቅጣጫ 500 ሜትር አለፍ ብሎ በግራ በኩል ጀነራል መርካንታይል ህንፃ አጠገብ በሚገኘው የድርጅቱ መ/ቤት በግንባር በመገኘት መመዝገብ የምትችሉ መሆኑን እንገልጻለን። የሥራ መደቡ መጠሪያ፡ሲኒየር ግራፊክስ ዲዛይነር የትምህርት ደረጃ፡ ዲፕሎማ በግራፊክስ ዲዛይነር/ፋይን አርት የሥራ ልምድ፡4 /አራት/ ዓመት በማስታወቂያ ሥራ ላይ የሠራ ብዛት፡1 /አንድ/የቅጥር ሁኔታ፡በቋሚነት ደመወዝ፡በስምምነት የስራ ቦታ፡አዲስ አበባ አድራሻ፡ ስ.ቁ. 0116 51 33 61

Appendix II: Sample selected test data

Appendix II: Sample selected candidate texts

ክፍት <N> የስራ <NP> መደብ <N> ማስታወቂያ <ORG> ቤቴል <ADJ> ቲቺንግ <ADJ> ጠቅላላ <ADJ> ሆስፒታል <N> </ORG> ከዚህ <PRONP> በታች <PREP> በተጠቀሰው <VP> ክፍት <N> የስራ <NP> መደብ <N> ብቁ <ADJ> ባለሙያዎችን <NUMP> አወዳድሮ <NUMP> ለመቅጠር <NUMP> ይፈልጋል <V> ። <PUNC> የስራ <NP> መደቡ <V> መጠሪያ <POS> የራጅ <PREP> ክፍል <N> ቴክኒሽያን <ADV> </POS> ጾታ <N> ሴወ <N> የትምህርት <QUL> <NP> ደረጃ <N> ዲፕሎማ <N> </QUL> ። <PUNC> ብዛት <N> 1 <NEED> የስራ <NP> ቦታ <N> አዲስ <NUMP> አበባ <ADJ> ደመወዝ <VN> በስምምነት <SAL> የቅጥር <NP> ሁኔታ <N> በቋሚነት <AGREEMENT> የስራ <NP> ልምድ <N> ከ2 <NUMP> አመት <NUMP> በላይ <PREP> የማመልከቻ <NP> አድራሻ <NC> ወይራ <N> ሰፈር <N> መስፈርቱን <VN> የምታሟሉ <VREL> አመልካቾች <NUMP> ይህ <PRON> ማስታወቂያ <N> ከወጣበት <VP> ቀን <N> ጀምሮ <V> ባሉት <VP> <DEAD> <NUMCR> ሰባት <NUMCR> ተከታታይ <ADJ> የስራ <NP> ቀናት <N> </DEAD> <NUMP> ቢሮ <N> ቁጥር <N> 114 <NUMCR> በመቅረብ <NP> ማመልከትና <NC> መመዝገብ <VN> የምትችሉ <NP> መሆኑን <VN> እንገልጻለን <NUMP> ። <PUNC> <ORG> ብርሀንና <NPC> ሰላም <N> ማተሚያ <N> ድርጅት <N> </ORG> ባለው <NUMP> ክፍት <N> የስራ <NP> መደብ <N> ላይ <PREP> መቅጠር <VN> ይፈልጋል <V> ። <PUNC> የስራ <NP> መደቡ <POS> ከፍተኛ <ADJ> የምርት <NP> እቅድና <NC> ክትትል <N> ባለሙያ <NUMP> </POS> ለ2ኛ <NUMP> ጊዜ <N> የወጣ <VREL> <ADJ> የትምህርት <QUL> <NP> ደረጃ <N> ከታወቀ <VP> ዩኒቨርስቲ <N> ኮሌጅ <N> በኢንዱስትሪያል <ADJ> ምህንድስና <NC> በህትመት <NP> ቴክኖሎጂ <N> በኢንዱስትሪያል <ADJ> ቴክኖሎጂ <N> በማኑፋክቸሪንግ <NP> ቴክኖሎጂ <N> የመጀመሪያ <NP> ዲግሪ </QUL> <N> ። <PUNC> 4 <NUMCR> አመት <NUMP> አግባብ <NUMP> ያለው <NUMP> የስራ <NP> ልምድና <NC> የህትመት <NP> ሙያ <N> ከህትመት <NP> ቴክኖሎጂ <N> ተመራቂዎች <N> በስተቀር <PREP> የኮምፒዩተር <NP> ስልጠና <N> የወሰደ <VREL> ደረጃ <N> 16 <NUMCR> ደመወዝ <VN> 3572 <SAL> ብዛት <N> 1 <NEED> አንድ <NUMP> የስራ <NP> ቦታ <N> አዲስ <NUMP> አባባ <NC> የቅጥር <NP> ሁኔታ <N> በቋሚነት <AGREEMENT> የስራ <NP> ልምድ <N> ከምረቃ <ADJ> በኋላ <PREP> ሲሆን <VP> መስፈርቱን <VN> የሚያሟሉ <VREL> አመልካቾች <NUMP> ይህ <PRON> ማስታወቂያ <N> ከወጣበት <VP> ቀን <N> አንስቶ <NUMP> <DEAD> <ADJ> 10 <NUMCR> ተከታታይ <ADJ> የስራ <NP> ቀናት <N> </DEAD> <PREP> የትምህርትና <NPC> የስራ <NP> ልምድ <N> ማስረጃ <N> ዋናውንና <ADJ> የማይመለስ <NUMP> ፎቶ <N> ኮፒና <NC> ጭV <CONJ> ከማመልከቻ <NP> ጋር <PREP> በማያያዝ <NP> በድርጅቱ <NP> 7ኛ <NUMOR> ፎቅ <N> ቢሮ <N> ቁጥር <N> 12 <NUMCR> የሰው <NP> ሀብት <N> ስራ <N> አመራር <NUMP> ቡድን <N> ማመልከት <VN> የሚችሉ <VREL> መሆኑን <VN> እንገልጻለን <NUMP> ። <PUNC> የቅጥር <NP> ሁኔታ <N> የለም <AGREEMENT> ። <PUNC> ክፍት <N> የስራ <NP> ቦታ <N> ማስታወቂያ <N> ድርጅታችን <ORG> ባማኮ <N> ኢንጅነሪንግ <NUMP> ሀላፊነቱ <N> የተወሰነ <VREL> የግል <NP> ማህበር <N> </ORG> ከዚህ <PRONP> በታች <PREP> ለሚመለከተው <NUMP> ክፍት <N> የስራ <NP> ቦታዎች <N> አወዳድሮ <NUMP> መቅጠር <VN> ይፈልጋል <V> ።የስራ <NC> የስራ <NP> መደቡ <V> መጠሪያ <POS> ጥበቃ <N> </POS> <ADJ> የትምህርትና <NPC> የስራ <NP> ልምድ <N> 4ኛ <NUMOR> ክፍል <N> የስራ <NP> ቦታ <N> አዲስ <NUMP> አበባ <ADJ> ብዛት <N> 10 <NEED> ደመወዝ <VN> 800 <SAL> በመሆኑም <NPC> አመልካቾች <NUMP> ይህ <PRON> ማስታወቂያ <N> በጋዜጣ <NPC> ከወጣበት <VP> ቀን <N> አንስቶ <NUMP> <DEAD> <VP> 5 <NUMCR> ተከታታይ <ADJ> የስራ <NP> ቀናት <N> </DEAD> <PREP> <NP> እና <NUMOR> የስራ <NP> ልምድ <N> ማስረጃችሁን <VN> በመያዝ <NP> የድርጅቱ <NP> ዋና <ADJ> መስሪያ <N> ቤት <N> በመምጣት <NP> መመዝገብ <VN> የምትችሉ <NP> መሆኑን <VN> እንገልጻለን ። <PUNC> ደመወዝ <VREL> ከ1000 <SAL> <NUMP> ብር <N> በላይ <PREP> ሆኖ <V> ከሚሰራበት <VP> መቤት <N> ሀጋዊ <ADJ> ደብዳቤ <N> ማቅረብ <VN> የሚችሉ <VREL> መሆን <VN> ይጠበቅበታል። <NUMP> አድራሻ <NC> ከቦሌ <NP> ጉምሩክ <N> ፊት <N> ለፊት <NUMP> ከሚገኘው <VP> አሚን <NUMP> ጀነራል <N> ሆስፒታል <N> ወደ <PREP> ታች <PREP> ዝቅ <N> ብሎ <V> በሚገኘው <VP> ቅያስ <N> ወደ <PREP> ውስጥ <PREP> 200 <NUMCR> ሜትር <N> ገባ <N> ብሎ <V> ይገኛል <V> ።ለበለጠ <NUMP> መረጃ <N> 0118956942 <PHONE> <NUMCR> የቅጥር <NP> ሁኔታ <N> የለም <AGREEMENT> ብዛት <N> የለም <NEED> ። <PUNC>

Appendix III

A. Sample code for data preprrocessing

```
def charnormal():
    doc=open("lasttesting.txt",'r', encoding = 'utf-8' )
    y=doc.read()
    simchar=open("similar char.txt",'r', encoding = 'utf-8' )
    yy = simchar.read().split()
    dic = {}
    for k in range(0,len(yy),2):
        dic[str(yy[k])] = str(yy[k+1])

    lis=[]
    yy=''
    lis=list(dic.keys())
    for i in y:
        for j in lis:
            if j==i:
                y = y.replace(i, dic[j])
    #print(y)
    return y
charnormal()
def wordnorml():
    z = charnormal()
    sp = z.split()
    simchar=open("similarWord.txt",'r', encoding = 'utf-8' )
    a = simchar.read().split()
    dic = {}
    for k in range(0,len(a),2):
        dic[str(a[k])] = str(a[k+1])
    lis=[]
    a = ''
    lis=list(dic.keys())
    for i in range(0,len(lis)):
        for j in range(0,len(sp)):
            if sp[j]==lis[i]:
                sp[j]=sp[j].replace(sp[j],dic[lis[i]])
    for i in range(0,len(sp)):
        a += sp[i] + ' '
    #print (a)
    return a
def punkremv():
    zz = wordnorml()
    y = re.sub('[,.!፡["()"+፧?/\]!-]','',zz)
    doc=open("lastpreprocessecd.txt",'w', encoding = 'utf-8' )
    doc.write(y)
```

```
            doc.close()
```

B. Sample code for candidate text selection

```
org = open('org gazeter.txt','r', encoding = 'utf-8')
orggaz = org.readlines()
x = open("position gazeeter.txt",'r', encoding = 'utf-8')
y = x.readlines()
xx = open("lastdecodendtagged.txt",'r', encoding = 'utf-8')
yy = xx.read().split()
ay=[]
xxx = ' '
for k in range(0,len(yy)):
    for i in range(0,len(y)):
        y[i]=y[i].rstrip('\n')
        a = y[i].split()
        if a[0] == yy[k]:
           if a[(len(a)-1)] == yy[(k+(len(a)-1))]:
              del yy[k-1]
              #del yy[k+(len(a)-1)]
              yy.insert(k-1,'<POS>')
              yy.insert((k+len(a))+1,'</POS>')
for j in range(0,len(orggaz)):
     orggaz[j]=orggaz[j].rstrip('\n')
     org1 = orggaz[j].split()
     for k in range(0,len(yy)):
         if org1[0] == yy[k]:
            if org1[(len(org1)-1)] == yy[(k+(len(org1)-1))]:
               del yy[k-1]
               yy.insert(k-1,'<ORG>')
               yy.insert((k+len(org1))+1,'</ORG>')
for i in yy:
    xxx += i + ' '
m = open('organization nd position tagging.txt', 'w', encoding = 'utf-8')
m.write(str(xxx))
m.close()
print("organization and job position is already tagged")

print("\n==========================================\n\n")

y = open('organization nd position tagging.txt', 'r', encoding = 'utf-8')
yy = y.read()
sp=yy.split(' ')
for i in sp:
    y = open('allcandidates.txt', 'w', encoding = 'utf-8')
    for i in range(len(sp)):
            if sp[i] == "ብዛት":
```

```
    del sp[i+3]
    sp.insert(i+3,'<NEED>')
if sp[i] == "የቅጥር":
    if sp[i+2] == "ሁኔታ" or sp[i+2] == "አይነት":
        del sp[i+5]
        sp.insert(i+5,'<AGREEMENT>')
if sp[i] == "ደመወዝ" or sp[i] =='ደሞዝ':
    del sp[i+3]
    sp.insert(i+3,'<SAL>')
if sp[i] == "አመት":
    if sp[i+2] == "ከዚያ":
        if sp[i+4] == "በላይ":
            del sp[i-1]
            del sp[i+1]
            sp.insert(i-2,'<EXPER>')
            sp.insert(i+1,'</EXPER>')
if sp[i] == "ተከታታይ":
    if sp[i+2] == "የስራ":
        if sp[i+4] == "ቀናት":
            del sp[i+6]
            del sp[i-4]
            sp.insert(i+5,'</DEAD>')
            sp.insert(i-4,'<DEAD>')
    elif sp[i] == "ተከታታይ" or sp[i] == "የስራ":
        if sp[i+2] == "ቀናት":
            del sp[i+4]
            del sp[i-3]
            sp.insert(i+3,'</DEAD>')
            sp.insert(i-3,'<DEAD>')
if sp[i] == "ስልክ":
    if sp[i+2] == "ቁጥር":
        del sp[i+5]
        sp.insert(i+5,'<PHONE>')
    elif sp[i] == "ስልክ" or sp[i] == 'መረጃ':
         del sp[i+3]
         sp.insert(i+3,'<PHONE>')
if sp[i] == "አመት":
    if sp[i+2] == "ከዚያ":
        if sp[i+4] == "በላይ":
            del sp[i-3]
            del sp[i+2]
            sp.insert(i-3,'<EXPER>')
            sp.insert(i+2,'</EXPER>')
    elif sp[i] == "አመት":
        if sp[i+2] == "የስራ":
            if sp[i+4] == "ልምድ" or sp[i+4] =='ልምድና':
```

```
                    del sp[i-3]
                    del sp[i+2]
                    sp.insert(i-3,'<EXPER>')
                    sp.insert(i+2,'</EXPER>')
                if sp[i] == "አመት":
                  if sp[i+2] == "የሰራ":
                    del sp[i-3]
                    del sp[i+1]
                    sp.insert(i-3,'<EXPER>')
                    sp.insert(i+2,'</EXPER>')
            if sp[i]=="የትምህርት" or sp[i]=="ተፈላጊ" :
              if sp[i+2] =="ደረጃ" or sp[i+2]=="ችሎታ" :
                y.write(sp[i]+' ')
                y.write("<QUL>")
                y.write(' ')
            elif sp[i]=="የተመረቀች" and sp[i+3]=="<PUNC>" :
                y.write(sp[i])
                y.write(' ' + "</QUL>" + ' ')
            elif sp[i] == "ዲግሪ" and sp[i+3]=="<PUNC>":
                y.write(sp[i])
                y.write(' ' + "</QUL>" + ' ')
            elif sp[i] == "ዲፕሎማ" and sp[i+3]=="<PUNC>":
                y.write(sp[i])
                y.write(' ' + "</QUL>" + ' ')
            else:
                y.write(sp[i] + ' ')

#print(sp)
y.close()
print("All candidates are taggegs")
print("\n==========================================\n")
```

C. Python code for selected and tagged text extraction

```
outfile =codecs.open('lastttt.txt','w',encoding='utf-8')
y = codecs.open("allcandidates.txt", 'r', encoding = 'utf-8')
yy = y.read()
z=yy.split()
i=0
x=len(z)-1
y=''
m=''
while i<x:
    c=0
    if z[i]=="<ORG>":
        c=i+1
        y+='\n'
        while(z[c])!="</ORG>" :
            #a.append(z[c])
            y += ' '+ z[c]
            c+=1
            if c==x:
                break
        i=c+1
    if c==0:
        i+=1
    else:
        continue
v=0
x=len(z)-1
bb=''
while v<x:
    d=0
    if z[v]=="<POS>":
        d=v+1
        bb+='\n'
        while(z[d])!="</POS>" :
            bb += ' '+ z[d]
            #b.append(z[d])
            d+=1
            if d==x:
                break
        v=d+1
    if d==0:
        v+=1
    else:
        continue
vv=0
```

```
for jj in y:
    yyy +=  re.sub('[<PRONVCUMADJEL>]','',jj) + ''
for ii in bb:
    mm +=  re.sub('[<PRONVCUMADJELáˆˆá‹¨>]','',ii) + ''
for iiii in bbb:
    bbbb +=  re.sub('[<PRONVCUMADJEL>]','',iiii) + ''
for dd in dead2:
    dead +=  re.sub('[<PRONVCUMADJELʌዮ>]','',dd) + ''
for qq in qul:
    quli +=  re.sub('[<PRONVCUMADJEL>]','',qq) + ''
xx=[ ]
ll=[ ]
iii=[ ]
kk=[ ]
for w in range(len(z)):
    if z[w] == '<SAL>':
       ll.append(z[w-1])
for j in range(len(z)):
    if z[j] == '<NEED>':
       xx.append(z[j-1])
for u in range(len(z)):
    if z[u] == '<AGREEMENT>':
       iii.append(z[u-1])
for t in range(len(z)):
    if z[t] == '<PHONE>':
       kk.append(z[t-1])
a=yyy.split('\n')
m=mm.split('\n')
ex=bbbb.split('\n')
dead1=dead.split('\n')
qulif=quli.split('\n')
for i in range(0,len(a)):
     for j in range(0,len(m)):
         for exper in range(0,len(ex)):
           for n in range(0,len(dead1)):
              for qulification in range(0,len(qulif)):
                 for salary in range(0,len(ll)):
                    #for ned in range(0,len(xx)):
                       if i==j==exper==n==qulification==salary:#==ned:
                             cursor.execute("INSERT INTO Table_4
                                (ORG,POS,QUL,SAL,EXPER,DEAD)VALUES(?,?,?,?,?,?)",a[i],m[j],
                                qulif[qulification],ll[salary],ex[exper],dead1[n])
                          cursor.commit()
```

DECLARATION

I, the undersigned, declare that this project 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 project have been duly acknowledged.

_____

Sintayehu Hirpassa

June 2013

This thesis has been submitted for examination with my approval as university advisor.

_____

Ermiyas Abebe