ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

INFORMATION EXTRACTION MODEL FROM AMHARIC NEWS TEXTS

BY: Getasew Tsedalu

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

November, 2010

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF COMPUTER and MATHEMATICAL SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

INFORMATION EXTRACTION MODEL FROM AMHARIC NEWS TEXTS

BY: Getasew Tsedalu

ADVISOR:

Solomon Atnafu (PhD)

APPROVED BY
EXAMINING BOARD:
1. Dr. Solomon Atnafu, Advisor _____
2. _____ _____
3. _____ _____

Dedication

For dad: You have been working so hard to raise me, my two brothers and my sister. You are the hero in my life the things you did to me and our family will always stay in my heart. May God bless you and give you a long life and good health.

For mom: you are always positive and supportive of things that I do. I thank you for being the light of my life.

Acknowledgment

All the praise goes to God, who gives me the strength from beginning to the end of this research work. Doing something new has a lot of obstacles which sometimes make someone to quit instead of striving for its successful completion. In those times to overcome the struggle that I face different individuals has their own contribution for the successfully completion of the work. The first person I would like to thank is my advisor Dr. Solomon Atnafu for his suggestive and constructive comments which strengthen this research work. I would also like to acknowledge Tesfaye Guta, who has been working day in day out with me in the lab. I extend my thanks to the officials in ENA for providing me the necessary data that is important for training and testing of the prototype developed. Finally my gratitude goes to Abebe Abeshu, Teklay G/Egiziabher, Mequannt Munye, Mandefro Kejela, Esmael Kedir, Desta Berihu, Michael Shiferaw, Moges Ahmed and all my classmates for the discussion we have and for the ideas we share which was very helpful for the successful completion of this work.

Table of Contents

List of Tables

List of Figures

Acronyms & Abbreviations

IE: Information Extraction

IR: Information Retrieval

ATIE: Amharic Text Information Extraction

NLP: Natural Language Processing

MUC**:** Message Understanding Conference

ARFF: Attribute Relation File Format

GATE: General Architecture for Text Engineering

NLQA: Natural Language Question Answering

ST: Scenario Template Production

TR: Template Relation Construction

TE: Template Element Construction

NER: Named Entity Recognition

POS: Part of Speech Tagger

ENA: Ethiopian News Agency

ROC: Receiver Operating Characteristic

Abstract

As the growth of unstructured documents in the web and intranet is increasing from time to time, a tool that can extract relevant data to facilitate decision making is becoming crucial. IE is concerned with extraction of relevant information from text and stores them in a database for easy use and management of the data. As the first comprehensive work on IE from Amharic text we designed a model that is genuine enough to deal with different domains in the Amharic language. The proposed model has document preprocessing, text categorization, learning and extraction and post processing as its main components. The document preprocessing component handles the normalization of the document while text categorization and learning and extraction handle the categorization of the news text and extracting the predefined relevant information from the categorized text respectively. The post processing component format and save the extracted data to the database.

Various evaluation techniques, which are used to evaluate the performance of the classifier machine learning algorithms, are used for IE and text categorization. Among the different classifier machine learning algorithms used for text categorization component, the Naïve Bayes algorithm performs by correctly classifying 92.83% of the 1200 news texts used as a dataset. On the other hand, 1422 instances are used for training and testing the Information Extraction component. Different scenarios are used to evaluate the role of the different features in predicting the category for the candidate texts. Among the different scenarios we considered and the different machine learning algorithms we employed the SMO algorithm correctly classified 94.58% of the instances correctly, when all the features are considered which yields higher precision and recall rate for the different attributes considered for extraction.

Key words: Amharic Text Information extraction, Machine Learning Approach to Information Extraction, Amharic Text Categorization, Information Extraction

CHAPTER ONE

1. INTRODUCTION

1.1 General Background

The explosive growth of documents on the web and intranet is highly increasing the amount of information available in unstructured and semi structured format from time to time. Every year a large amount of books, journals and other materials are published which will take more than a life time for human beings to read and use the important information contained in them [1]. The numbers of Amharic documents on the web and in other machine readable forms are also increasing from time to time. As a result of this growth, the huge amount of text which contains different valuable information which can be used in education, business, security and other many areas are hidden under the unstructured representation of the textual data. This shows that getting the right information for decision making from existing abundant unstructured text is a big challenge. The non availability of tools for extracting and exploiting the valuable information which is effective enough to satisfy the users need have been a major problem for years [2].

Information analysts, who have been working in information extraction for the database input, have to do the extraction manually. They have to go through all the text written in the document and extract the information that will be used as an input for the database, categorize them according to their properties and identify their relationship manually. Extracting the information manually from a large amount of unstructured text is a very tiresome and time consuming job [3]. On the other hand, 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.

To solve this problem a lot of researches have been conducted in the area of information extraction, information retrieval and NLP. Information Extraction (IE) is the automatic extraction of facts from text, which includes detection of named entities, entity relations and

events used to extract facts from unstructured text. It is based on analyzing natural language in order to extract information. The process takes texts (and sometimes speech) as an input and produces structured data as an output. This data may be used directly by users, or may be stored in a database or spreadsheet for later analysis, or may be used for indexing purposes in information retrieval (IR) applications such as web search engines [4, 5].

Information extraction is quite different from information retrieval. An IR system finds relevant texts that is based on a query and presents them to the user while an IE application analyzes texts and presents only the specific information from it 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 query statements and receive in return a set of documents (e.g., newspaper articles) that contain likely matches. The User then would have to read the documents and extract the necessary information from the retrieved data. In contrast, an IE system would automatically populate a spreadsheet directly with the names of relevant information instead of the text which contain the information. There are advantages and disadvantages with IE in comparison to IR. IE systems are more difficult and knowledge-intensive to build, and are to varying degrees tied to particular domains and scenarios. IE is more computationally intensive than IR. However, 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].

Information extraction has three different components regardless of the language and domain on which it is developed for, which are linguistic preprocessing, the learning and application stage and semantic post processing to do the extraction of data from a given text. The 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 and the last component which is the semantic post processing assign the extracted information into their predefined attribute category and manages the normalization and duplication problem with the extracted data [6].

Information extraction unlike the other research domains is language and domain dependent [7]. Different researchers tried to develop information extraction systems for different languages and

domains. The IE system developed for English and any specific domain in English cannot work for Amharic language as there are different language and domain specific issues which cannot be handled by the system developed for English. This makes it there is a need to do a research in extracting data from the Amharic texts as most of the information that exist in Amharic language is unstructured. This initiates us to engage in research to develop a model for extracting facts from the Amharic news texts and then develop a prototype to realize the efficiency of the model in extracting facts which will be used as an input for the database system.

1.2. Statement of the Problem

The IE tools developed so far are language and domain dependent. The information extraction system developed for English or any other language and for some specific domain cannot work for other languages of the same domain. 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. If someone try to use an information extraction system which is developed for English and use it to work on the same domain for Amharic the result will be very poor. Thus, in this work we will design an effective information extraction model for Amharic language as to the best knowledge of the researcher there is no information extraction system which is developed for Amharic language so far.

1.3 Objective

General Objective

The general objective of this research work is to design a generic model of information extraction system from Amharic news text using a classification machine learning approach and evaluate its efficiency by developing a prototype Amharic text IE System.

Specific Objective

To achieve the general objective, the following specific objectives are identified:

- Analyse the general grammatical structure of Amharic language for the purpose of identifying its characteristics for IE.

- Identify the representation of different entities on Amharic news text and its grammatical structure.
- Develop a model for IE from Amharic news text
- Use the existing algorithms or developing a new algorithm for IE from Amharic news text
- Identify and collect a corpus of the Amharic news text.
- Develop a prototype system based on the model
- Evaluate the effectiveness and usability of the algorithm developed for Amharic IE

1.4 Scope and Limitation

Information extraction is a very complex and rigorous task which needs the understanding of the natural language and the specific domain in need. A full-fledged information extraction system will require a number of Natural language processing 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 some researchers, they are not publicly available. Having these limitations in mind, the scope of the research work is limited in extracting numeric and name data from the Amharic news text and assign them to the predefined database slot. The 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.

1.5. Methodology

To achieve the general and specific objectives of the study, the following methods are employed

1.5.1 Literature Review

Literature review have been conducted on books and research works which are related to information extraction and related fields to have understanding of the problem domain and the related researches that are done so far, which helps for the better understanding to design a new model for Amharic language.

1.5.2 Analysis of the Amharic Language

The different facts about Amharic language like the grammatical structure the number representation and other language specific issues that are important for the research work have

been reviewed and presented. It helps to understand the nature of the language with regard to information extraction.

1.5.3 Design and Implementation of ATIE System

A classification machine learning approach is used to develop ATIE (Amharic Text Information Extraction) system. It contains the document preprocessing, text categorization, extraction and learning and post processing as a four main components.

1.5.4 Development Tools

In order to develop a prototype system, different appropriate tools have been selected and used. The java programming language is used to implement the different language specific algorithms developed. Weka is used to handle the classification of the candidate text and news text in case of information extraction and text categorization respectively. The POS which is used as one of the features in IE component is developed using HMM (Hidden Markov Model) available in the Ling Pipe open source NLP software.

1.6 Application of Results

IE system for Amharic news text can be used to extract information that exists in the news text in the way that it can be used as an input for database or for other applications. Nowadays most of the available information is in an unstructured textual format. Since we cannot query it in simple ways, automatic systems cannot use the contained information and human beings cannot easily access and manage it. The IE system for Amharic news text can be used for better use of the contained information in the Amharic news texts.

Implementing the IE system for the Amharic news text will have the following uses:

I. The unstructured information which exist in the Amharic news text will be structured, which make it easy for the different users to access it within a short period of time. For example, if somebody wants to know the infrastructures that are developed in different part of the country, he/she doesnot have to read all the news papers. The IE system for the Amharic news will extract the summarized information related to infrastructure and store it in the database for easy use and management. The users can then easily access the data by using a database query language.

II. The extracted structured information from unstructured text can be used as an input for the other applications. Since that data is structured, applications can share the information and use it for different purposes.

1.7 Thesis Organization

The rest of this thesis report is organized as follows. Chapter 2 discusses the different issues in IE and the related subject areas as literature review. This Chapter lays the foundation in understanding what an IE system comprises of, what approaches are used, and the different components which are needed by IE system. Chapter 3 is devoted to discuss related works done on IE systems in different languages and on different domains. Chapter 4 discusses the Amharic language with regard to IE. Many language specific issues such as the writing system have been presented. The architectural and design issues of our system are discussed in Chapter 5. The main components of our system, their functional operation and the specific sub-component of each component are briefly discussed in this Chapter. Chapter 6 is devoted to discussing the main implementation issues of our IE system. The algorithms, techniques and methods used in how the system has been successfully developed are discussed in this Chapter. Chapter 7 presents the evaluation of the system. Chapter 8 concludes the thesis by outlining the contribution of the research work. It also shows some research directions that can be used in the future to improve IE system for Amharic.

CHAPTER TWO

2. LITERATURE REVIEW

In this Chapter, we provide a brief overview of the field of information extraction. The different components of IE and the approaches used to develop information extraction systems are reviewed and presented. The related NLP fields to information extraction 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.

2.1 Information Extraction

Now a days, digitally stored information is available in abundance and in a numerous forms to an extent of making it near impossible to manually search, filter and choose which information one should use for his/her own purpose. This abundance of information must instead be filtered and extracted in order to avoid drowning in it [9]. Different scholars in the area tried to develop different information management systems so that the drowning 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 problem which solves it in different ways are Information Retrieval (IR), Information Extraction (IE), Question and Answering, Text Summarization, Text Categorization, etc.

Information Extraction (IE) addresses the problem of information overload by locating the target phrases from document and transforms them in to structured representation. As it is defined by MUC [10], IE is the "task of extraction of information from a text in the form of text string which are placed in to slots labeled to indicate the kind of information that can fill them". Given a free text an IE system will extract the specific information from the text and put them in the database so that they can easily be retrieved and managed [11].

Un Yong Nahm [12] also defines IE as "a form of shallow text understanding that locates specific pieces of data in natural language documents, transforming unstructured text into a structured database". IE is all about extracting structural factual data mostly from unstructured

text (web pages, text documents, office documents, presentations, and so on). IE usually uses NLP tools, lexical resources and semantic constraints for better efficiency. The General Architecture for Text Engineering (GATE) [13], which is the widely known open source software system for computations related to natural language, defines Information Extraction (IE) as a system which analyses unstructured text in order to extract information about pre-specified types of events, entities or relationships.

2.2 Related NLP Fields to Information Extraction

2.2.1 Information Retrieval

Information Retrieval (IR), which is a relatively matured field of study, provides a method that is being widely used to acquire a set of relevant material from a huge amount of data collection especially in unstructured form. It is defined as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)" [3]. IR system interface takes a natural language text from the user and convert it to equivalent query and retrieve the document from abundance of documents which contain the data that the user wants [3]. The number of resulting pages returned by the IR system could be very large and not all information contained in a single page is interesting, which requires further refinement by the user [11]. Even if the information retrieval highly minimizes the abundance of documents to small number of documents that are more appropriate to the users need they are still large enough for the user to use them for decision making.

2.2.1 Text Summarization

Text summarization (or rather, automatic text summarization) is the technique where a computer automatically creates an abstract, or summary, of one or more texts. The initial interest in automatic text summarization started during the nineteen sixties in American research libraries. A large amount of scientific papers and books were to be digitally stored and made searchable. However, the storage capacity was very limited and full papers and books could not be fit into databases those days. Therefore, summaries were stored, indexed and made searchable [9]. Radev [14] defines a summary as "a text that is produced from one or more texts, that convey important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that". The summary can be created from a single or multiple

documents by applying different methodologies. This simple definition captures three important aspects that characterize research on automatic summarization

- Summaries may be produced from a single document or multiple documents,
- Summaries should preserve important information and
- Summaries should be short

Even if the text summarization minimize the large amount of text by extracting the most relevant sentence from larger text it is still requires the user involvement to read the summary and extract the specific information she/he needs and the data is not yet easily used by other computer application directly.

2.2.3 Question and Answering System

Natural Language Question answering (NLQA) retrieves a textual fragment from a document collection that represents the answer to a question. NLQA extract the specific information from the given document which is retrieved by using IR system based on the given question by the user. It might return one or a set of textual fragments from a document collection that represents the answer to the question. The output of NLQA is unstructured information based on the given question by the user [15]. NLQA is profoundly different from Information Retrieval (IR) or Information Extraction (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 NLQA 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.

The different solutions that are listed above to solve the information overload issues have their own advantage and disadvantage in solving the problem. IE unlike text summarization, IR and

question and answering system its output is in structured form which can be easily managed and accessed as it is stored in the database. Different students at the postgraduate level try to develop a question and answering, text summarization system for Amharic language which will help to easily manage the information processing and use of the Amharic language in the computer system. IE system will rather simplify the problem more by supporting the extraction of the specific information the user wants from the given set of Amharic news document and fill the predefined slots attributes and the output can be used by other application for different purpose directly unlike NLQA, text summarization and information retrieval.

2.2.4 Text Categorization

Text categorization is a process involving the assignment of predefined categories to free (unlabeled) text data. With the rapid growth of online data and information, text categorization has become one of the key techniques for processing and organizing text documents. Text categorization techniques are used to classify new stories, to find interesting information on the World Wide Web and to guide a user's search through hypertext. The output of text categorization is a category of the text [33].

2.3 Architecture of IE System

Different IE systems for different languages and different domains using different approaches are developed so far and are still on development but they all use the different task breakdown for IE. The research work in [38] categorizes IE in to five different tasks.

- Named Entity Recognition
- Co reference resolution
- Template Element Construction
- Template Relation Construction
- Scenario Template Production

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 [16].

For example the NER takes the following Amharic news text obtained from ENA and recognize the named entities and numbers which will be used as attributes for the predefined database slot
ባሌ ዞን በመገባደድ ላይ ባለው የበጀት ዓመት 25 ሚሊዮን ብር በላይ ካፒታል ላስመዝግቡ ባለሃብቶች የኢንቨስተመንት ፈቃድ መሰጠቱን የዞኑ ኢንቨስትመንት ጽህፈት ቤት አስታወቀ፡፡ የጽህፈት ቤቱ ተወካይ አቶ ሁሴን መሐመድ ዛሬ እንደገለጹት በዓመቱ ፈቃዱ የተሰጠው በግብርናው ዘርፍ በፍራፍሬና በሰብል ልማት እንዲሁም በሆቴልና ቱሪዝም መስኮች ለተሰማሩ 15 ባለሃብቶች ነው፡፡ ባለሃብቶቹ ከዞኑ አስተዳደር ከ1 ሺህ 150 ሄክታር በላይ ቦታ መረከባቸውን ገልጸው፣ ከመካከላቸው አስሩ አስፈላጊውን ዝግጅት አጠናቀው ወደ ሥራ መግባታቸውንም ተናግረዋል፡፡ እንደ ተወካዩ ማብራሪያ እስካሁን ወደ ስራ የገቡት እነዚሁ ባለሃብቶች 1ሺህ ለሚሆኑ ወገኖች ቋሚና ጊዜአዊ የሥራ ዕድል ፈጥረዋል፡፡

The words which are names and numbers that represent different thing will be extracted like ባሌ, 25, ኢንቨስትመንት ጽህፈት ቤት, ሁሴን መሐመድ, 15, 1 ሺህ 150, 1ሺህ which are the named entity attributed in the text which represent different things.

Co-Reference Resolution

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 [39].

Template Element Construction (TE)

It is one of the IE task which associate descriptive information for the NER results. The different recognized name entities in association with their co reference are the input of template element construction [37]. The different recognized named entities will have different attributes For example, if we consider one of the entity organizations we might have the attribute like

Type: Which define the type of the organization whether it is government, commercial, non-governmental, etc

Name: name of the organization

Aliases: another form of representation of the organization name

From the example in the NER, the template element construction assign attributes to the recognized named entities the attributes might be represented as follows

Place: ባሌ

Amount of money for investment: 25 ሚሊዮን

Organization name: ኢንቨስትመንት ጽህፈት ቤት

Manager: ሁሴን ማሕመድ

No of investor: 15

Land in hectare: 1 ሺህ 150

No of employees: 1ሺህ

Template Relation Construction (TR)

Template relation task is the identification of possible relationship between template elements identified during the template element construction task. This might be, for example, an employee relationship between a person and a company, a family relationship between two persons, or a subsidiary relationship between two companies and other different relationship that exist in natural language text. Extraction of relations among entities is a central feature of almost any information extraction task, although the possibilities in real-world extraction tasks are endless [38].

Scenario Template Production (ST)

It is a combined output of template element construction and template relation construction and creates a scenario. Scenario templates (STs) are the prototypical outputs of IE systems, being the original task for which the term was coined. They tie together TE entities and TR relations into event descriptions. For example, TE may have identified Isabelle, Dominique and Françoise as People entities present in the Robert edition of Napoleon's love letters. ST might then identify Facts such as that Isabelle moved to Paris in August 1802 from Lyon to be nearer to the little chap, that Dominique then burnt down Isabelle's apartment block and that Françoise ran off with one of Gerard Depardieu's ancestors [38].

In another work survey conducted on information extraction the tasks are categorized in to five different and independent components [39]. As IE activity can be a very complex task decomposing it into different task is advantageous. The advantages of decomposition are that:

- It will be easy to choose the techniques and algorithms that suit each task,
- It will be easy to locally debug an Information Extraction program since the module responsible for each task is completely independent from the others and
- Information Extraction can be a customized activity according to an application's needs, by reordering, selecting and composing some of the tasks.

The considered tasks in the survey are:

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

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 whitespaces separating words, there are some cases in which simple whitespace 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 whitespaces 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. Lexicons and grammars can also be used in order to accomplish the task of segmentation using syntactic or lexical analysis. Another approach for segmentation of Chinese texts uses techniques based on statistics [39].

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 [39].

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 [39].

Normalization and Co-reference Resolution

Normalization and Co-reference resolution are the less generic tasks of the Information Extraction 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 [39].

The description of IE components presented in the above two works are almost the same even if they are presented using different names. The idea is that as IE is a complex task dividing it in to sub task will minimize the complexity of developing the system for IE. All the above listed components of IE can be categorized in to Linguistic preprocessing, learning and application phase and semantic post processing as the three main blocks regardless of the language, the domain and the approach on which the IE is developed for as presented in [2]. Each of the three blocks comprises tasks that are relevant for the selected information extraction approach. Even if the extraction algorithm and procedure are different in different approaches the basic issue behind all of them is to take text corpus and target structure as an input and produce structured data as an output by filling the predefined slots in the target structure [2, 6, 7].

2.3.1 Preprocessing of Input Texts

Text corpus often consists of unstructured, "raw" natural language texts. A big part of the relevant information can be distinguished by some regularity found in the linguistic properties of texts. In this phase the language property i.e. its structure, the position where most relevant information in the text are located, how the co-reference between sentences is presented in the text and other language and domain specific information will be studied and implemented into different linguistic components as part of the extraction system. The following linguistic components are proved to be useful for information extraction as they are described in [2].

Tokenization: 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 [8]. Tokenization is similar to segmentation presented in [38].

POS: Certain facts are typically expressed by certain parts of speech like names and determining parts of speech of tokens, known as POS tagging. Statistical systems can use POS tags as classification features, knowledge based systems as elements of extraction rules [2, 15].

(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.

2.3.2 Learning and Application of the Extraction Model

The learning and application phase gives much emphasis on the development of an IE system which can be adopted to the different domain and language easily which is not yet successful as IE is much bound to the language and the domain it is developed. Due to this reason the adaptation effort to the other domain is high. Modern IE systems use a learning component to reduce the dependence on specific domains and to decrease the amount of resources provided by human. An extraction model is defined according to the pursued approach and its parameters are optimized by a learning procedure. The three categories of approaches for IE use the learning method. 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 [16, 17, 18].

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 [16].

2.3.3 Post Processing of Output

The main motivation for IE is the structured representation of information that enables formal queries and automatic processing. One of the possibilities to structure the extracted data is to model the target structure as a database relation. 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 (e.g. representation of dates and numbers). 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 [16].

2.4 Approaches for Information Extraction

The approaches for information extraction can be categorized under two primary categories

- Knowledge engineering approach and
- Machine learning approach

These two 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. Earlier when IE was introduced the main focus was to extract the information from a short free text but latter its application is expanded to include the semi structured and structured texts [19].

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 [19, 20].

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 [19, 20].

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 [19, 20].

2.4.1 Knowledge Engineering Approach

Knowledge engineering approach relies on knowledge engineers i.e. human experts, who are skilled not only in the particular information extraction system, but also in the domain on which the IE system is developed for. In this approach the person has to read a set of domain relevant documents, find different rules by his observation and construct those rules by hand. In addition to requiring skill and detailed knowledge on a particular IE system, the knowledge engineering approach usually requires a lot of labor, a lengthy test-and-debug cycle, and it is dependent on having linguistic resources at hand, such as appropriate lexicons, as well as someone with the time, inclination, and ability to write rules. If any of these factors are missing, then the knowledge engineering approach becomes problematic. 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. The performance of the system is dependent on the competence of the knowledge engineer [19, 20, 21].

The advantage of knowledge engineering approach is good performing system is not hard to build as long as the knowledge expert is there and best performing system can be achieved by handcrafted rules. Its disadvantage is the development process is very laborious, some changes in specification can be hard to accommodate and required expertise that may not be available [22]. Therefore developing an IE system will be very difficult as there are non experts that know IE and the Amharic language and the NLP resources available in Amharic language are almost none.

2.4.2 Machine Learning Approach

The Machine learning approach is quite different from knowledge engineering approach. In 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. It is only necessary 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 IE system [11]. The machine learning approach for IE is then focused on the automatic acquisition of the extraction patterns that will be used to extract the information relevant to the particular task from each single document. When applying machine learning to IE, a learning

algorithm usually learns a model from a set of examples, grouped in documents, which have been manually annotated by the user. Then the model can be used to extract information from new documents. The accuracy of the learned model usually increases with the number of training examples made available to the system.

The strengths and weaknesses of the machine learning approach are complementary to those of the knowledge engineering approach. Rather than focusing on producing rules, the automatic training approach focuses on the training data. Corpus probabilities or rules are then derived automatically from the training data, and used to process novel or unseen data. As long as someone familiar with the domain is available to annotate texts, systems can be customized to a specific domain without intervention from any developers. The disadvantages of the automatic training approach also revolve around the fact that it is based on training data. Training data may be in short supply, or difficult and expensive to obtain. Sometimes one may wish to develop an extraction system for a topic for which there are few relevant examples in a training corpus. Such situations place a premium on the human intuition of a good rule designer. If the relations that are sought are complex or technical, if it is hard to find annotators, and if it is difficult to produce enough annotated data for training corpus then using machine learning approach will be disadvantageous [21].

Current IE approaches supported on supervised machine learning technique are divided in to the following there categories [23]

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, AutoSlog-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. Our method is related to the SRV system in that it models the IE task as a

classification problem. However, it applies Inductive Logic Programming and uses information about negative examples [23].

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 [23].

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. Their main advantage is that a deep linguistic analysis is not necessary instead classification techniques are used to find the desired information [23].

The method used in this research work is similar to all these systems. It is based on the same hypothesis. Looking around the words combination around the fact need to be extracted from the Amharic text is enough to know what type of information it is.

III Statistical Learning

This approach is focused on learning Hidden Markov Models (HMMs) as useful knowledge to extract relevant fragments from documents [23].

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. For instance, [19] compares three commonly used methods for named entity recognition SVM with quadratic kernel, maximal entropy, and a rule based learning system, and shows that the SVM-based system outperforms the other two. More features usually result in better performance and therefore, it is important to use the different features on the while using SVM algorithm to get a good performance.

2.5 Information Extraction as a Text categorization Task

IE is a token classification task rather than a Text categorization task. With IE we are working with texts but the basic unit that we are seeking to classify is tokens in the text rather than the entire text. With Text categorization we are seeking to identify whether an entire text is a member of particular category. With IE the categories are start and end, and the objects we seek to assign to these categories are the individual tokens. With IE we are representing individual tokens. We much encode additional information about the token to enable our learning algorithm to generalize. For IE we encode several features of the token as well as relational information about the surrounding tokens. The features include the specific token, as well as part-of-speech (POS), chunking, orthographic and gazetteer information. In addition, we add features to represent a fixed-width window of tokens on either side of the instance's token. The learning algorithm uses these features to create a model that can distinguish between tokens that are starts of fields, ends of fields or neither [24].

Let us first consider a two class classification problem. Let {(x1, y1)… (xn, yn)} be a training data set, in which xi denotes an instance (a feature vector) and yi $\in\{-1,+1\}$ denotes a classification label. A classification model usually consists of two stages: learning and prediction. In learning, one attempt to find a model from the labeled data that can separate the training data, while in prediction the learned model is used to identify whether an unlabeled instance should be classified as +1 or -1. (In some cases, the prediction results may be numeric values, e.g. ranging from 0 to 1. Then an instance can be classified using some rules, e.g. classified as +1 when the prediction value is larger than 0.5.) [25].

2.6 Evaluation Metrics

Since the classifier algorithms are employed as an algorithm applied for text categorization and information extraction the standard evaluation techniques used for evaluating classifier algorithm are presented. The accuracy of a classifier is calculated by the classifiers' performance on the test data set. The prediction of the category by the classifier can be evaluated using a confusion matrix, which represents how well a classifier recognizes instances of different classes. For example, if we have Politics news and economy news as two classes for prediction and if there are 100 news instances for each category of the following confusion matrix table, The first row of the confusion matrix table indicates that among the 100 instances of Politics news texts the classifier correctly predict the category of 90 news instances correctly while 10 of the remaining news instances are incorrectly categorized as economy news. The same works for the second row which the classifier predicts the category of 96 economy news correctly while 4 of them are categorized incorrectly as politics news.

Table 2.1: Two Class Confusion Matrix Predicted Class

Politics news	Economy news	classes
90	10	Politics news
4	96	Economy news

In addition to accuracy the experimental dataset may also require other performance measures to analyze the experimental result in detail like sensitivity, specificity, recall, precision, F-measure, and ROC Area. Sensitivity is also referred to as the *true positive rate*, while specificity is the *true negative rate* (that is, the proportion of negative instances that are correctly identified).

Sensitivity or true positive rate is the proportion of positive instances that are correctly identified by the classifier and it is calculated as

$$Sensitivity = \frac{t\_pos}{pos} \qquad (2.1)$$

Where *t_pos* is true positives and *pos* is the number of positive documents

Specificity is the proportion of false positive instances that are correctly wrongly classified as another class by the classifier

$$Specificity = \frac{f\_pos}{pos} \qquad (2.2)$$

Where f_pos is false positive and *pos* is the number of positive documents

The recall, precision and ROC area are then calculated from the result of sensitivity or True positive rate and specificity or false positive rate.

Precision is the proportion of instances that are correctly classified which a true positive instance

$$Precision = \frac{t_{pos}}{t_{pos} + f_{pos}} \qquad (2.3)$$

Where *t-pos* is the number of true positives and *f-pos* is the number of false positives

Recall is the proportion of instances which are classified correctly over the total number of instances in the test dataset [44]

$$Recall = \frac{t_{pos}}{total\ number\ of\ instances} \qquad (2.4)$$

Where t-pos is number of true positive instances

F-Measure is an optimization criterion which often is used for tuning the threshold in binary decision, which is defined in terms of precision and recall [44].

$$F\beta\ (\text{precision}, \text{recall}\ ) = \frac{(\beta^2\ + 1)(precision\ \text{X}\ recall)}{\beta^2\ (\text{precision}) +\ \text{recall}} \ldots\ldots\ldots\ldots\ldots\ldots 2.5$$

Where β is the parameter allowing differential weighting of precision and recall

If recall and precision are given equal weight β=1, Equation (2.5) will be simplified to

$$F - measure = \frac{2(recall \text{ X } precision)}{(recall \ + precision\ )} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.6$$

ROC (Receiver Operating Characteristics) curves show the trade-off between the true positive rate (sensitivity) and the false positive rate for a given model. The area under the ROC curve is a measure of the accuracy of the model. If there a counter false positive for every true positive, the ROC curve will be diagonal line with a ROC value 0.5. As the number of true positive instances of the class increases and the false positive instances decrease the ROC value increase from 0.5 to 1. Is the ROC value is 1 the classifier is perfect with the all the instances are true positive [43].

CHAPTER THREE

3. RELATED WORKS

The extraction of information from a free text for constructing databases is a recent technology. Scholars have conducted IE researches using different approaches on different domains and languages. Among these research works, some of the works that are more relevant to our research, which are done using a machine learning approach are reviewed and presented in this chapter.

3.1 Information Extraction from English Text

Most of the research works in the area of information extraction are conducted in different domain on the English language. Among these research works the research work in [26] presents a text categorization approach to extract information from business cards and automatically process change of addresses from email messages. They scan 505 business cards to create a text file for training and testing the system which is then annotated manually. After scanning some cleaning operations are done to make sure that the text file is ready for being processed by the text classifier to extract information. They use the Naïve Bayes algorithm to categorize the different line of the text into Address, Name, Title, Company name, company Logo, Phone number (either voice or fax), Web address, Email address, Telex number, cable number, and other miscellaneous text from the scanned text. They claim that Naïve Bayes algorithm is reasonably accurate but it ignores important structural constraints like how the business card is arranged what attribute come after the other and so on. Such constraints are handled by HMM in their work.

Another work presented in [27] focuses on using double classification approach to build a text classifier that can extract information from the software job corpus. They review the different proposed novel techniques that were used before for IE. They argued that IE is domain dependent because it is based on an extensive syntactic and semantic analysis of the language and the documents under consideration. However, this way the resulting IE system are less applicable in domain where no semantic information is available or for corpora in other

languages than English, for which fewer syntactic processing tools exist. By considering these issues in the IE research they propose a solution that can solve the problem which is an intuitive approach consisting of two classification based machine learning loops. In the first loop they use a text classifier to select the sentence that might contain relevant information from the document. In the second loop they perform a deeper analysis of those relevant sentences by performing word level classification. They use a Naïve Bayes classification for the first loop and rule based classification for the second loop. They use POS tags for syntactic information to build rules for the second loop. No semantic information is used. Their approaches is tested on software job corpus and obtained good results on most of the template slot, although for some slots (language and area) the word classification approach is not yet capable of finding good rules for the others they got 77% recall and 65% precision.

Another work presented in [28] uses the SVM and different feature sets to build a text classifier for information extraction. They review all the approaches and different algorithms that are used so far in different IE researches. They use the approach that treats the identification of fragment text to extract start and end position as distinct token classification task. The instances are all the tokens in the document. All the tokens that begin a labeled field are positive instance for the start classifier, while all the other tokens become negative instances for this classifier. Similarly, the positive examples for the end classifier are the last tokens for each labeled field and the other instances are negative examples. Each instance has a set of features that describe the given token. The features used in the paper include the specific token as well as POS, chunking orthographic and gazetteer information. They present the ELIE algorithm which has two distinct phases. In the first phase ELIE simply learns to detect the start and end of the fragment to be extracted. Their experiment shows that the first phase generally has high precision but low recall. The second phase is designed to increase the recall. They find that very often false negatives are “almost” extracted (the start but not the end is correctly identified, or the end but not the start). In the second phase, which they call convergent boundary classification; ELIE is trained to detect either the end of a fragment given its beginning, or the beginning of a fragment given its end. In this way, ELIE iteratively converges to the correct boundary classification that is missed in the first phase. They evaluate their approach on the seminar announcement dataset and the job posting dataset.

Another work presented in [29] uses a hybrid approach to information extraction. 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 naïve 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.

3.2 Information Extraction from Portuguese Text

A research work presented in [30] uses rule learning approach which applies rule learning without ignoring stop words for learning text extraction rules from Portuguese text. The main aim of the research work presented was to consider a collection of Portuguese text documents in certain domain, annotate the text by identifying the element of interest from the text, and use the annotated document to generate training set which in turn is used to generate extraction rules. These rules can then be applied to new unseen documents.

The system presented in the paper use three phases for extraction of information
Phase 1: Data preparation – which is concerned with selecting the set of documents that are of interest for a particular IE task and identify relevant elements from those relevant documents.
Phase 2: Generating text extraction rules – used to generate extraction rules which the authors use the enhanced naïve Bayes method.
Phase 3: processing new documents- this phase is concerned with extracting relevant information from new documents by using the extraction rules obtained during training of the system.

The stop words like conjunction (e.g. "and", "or" etc), prepositions (with, without, from etc) or other similar categories are considered quite irrelevant have no effect on the final results. Therefore these stop words are removed during preprocessing in earlier works. Authors of this paper proved that stop words will have a contribution to improve the result and they proved this assertion by their experiment. Their approach has an F-measure of 80% after it is tested on Portuguese language on domain concerning House/Flat sales.

3.3 Information Extraction from Thai Text

The research work presented in [31] uses the knowledge engineering and the machine learning approach to extract information from Thai text. This work is different from the IE systems developed for English language as it has its own unique features because of the nature of the language structure in Thai text. Thai language doesn't have capital letters, words themselves are a combination of compound words, and there is no sentence demarcation and other language specific ambiguities in the structure of the text. This language ambiguity which is most common for Asian languages as the authors claim is difficult to apply the pattern or rule learning approach directly.

To solve the ambiguity in the language structure of Thai texts they use the knowledge engineering approach to solve the ambiguities and the machine learning approach to train the system for new things that is not in the knowledge. The system they design has the following different components

Word Tokenization: As the words used in Thai text are a combination of other words, the first step is to segment these words in to individual words. In this phase the Head Words Dictionary and rules for specific lexicon are used by the authors to improve the tokenization process.

Part of Speech Tagging: the next step is to tag the different part of the segmented text to its appropriate part of speech tag. For this phase also they use lexicon dictionary and training corpus to effectively tag the words in the text.

Surface Structure Analyses: Since there is no clear indication of sentence boundaries, it is not possible for the system to process the document using a sentence-based approach. In Thai text it is usually possible to identify paragraph breaks, taking a paragraph as one or more sentences separated by empty line. In Thai, a space may be used as a separation between contents, but it can also be used for other purposes, making the identification of contents unreliable and complicating the syntactic analysis. In this step the regions of information are identified using the spaces that separate between contexts. However, as there are fraudulent spaces, for instance the spaces between numbers, labels and abbreviations, to prevent fault identification they remove these spaces. To assist this phase they use Thai context free structure Grammar.

Extraction Engine: use a predefined concept definition and semantic class dictionary and the output of the above three phases as an input and extract the specific information from the text.

The authors tested their approach on the Thai import/export domain derived from ministry of commerce. The result of the experiment is a 42% precision and over 70% recall. The reason for the precision percentage to be low comes from the ambiguity of the sentence structure in which the parser is unable to determine the boundary of the sentences. They concluded that with efficient grammar parser for Thai and a word sense disambiguation module the precision of their system can be improved. However, the major problem with the system in its current form is that it requires substantial knowledge acquisition.

3.4 Information Extraction from Chinese Text

A work in [32] presents an IE from Chinese free text which has similar nature with Thai text as most of the Asian languages have similarities. 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.

[1] ttp://news.cina.com.cn/special/gov/index.shtml

3.5 Information Extraction from Spanish Text

Another research work presented in [23] uses machine learning approach for extracting information from natural disaster news on Spanish news papers. The authors present a machine learning approach in order to extract the information about characteristics and effects of natural disasters that will be used as a relevant resource for populating determined database. They select the machine learning approach as it is more appropriate to use when compared with knowledge engineering as they require a human expert in generating extraction patterns while machine learning is flexible by itself as it learn extraction patterns and as it can be easily trained to add modification.

The authors present a linear separator approach for learning patterns in the sentence of the text. This approach is based on the hypothesis that looking at the word combination around the relevant data is enough for learning the required extraction patterns. This creates an advantage as it doesn't require deep linguistic analysis of text and can be used for a more generalized purpose.

The proposed system presented in the paper is used to extract the following information from the disaster news. Information related to disaster itself such as Date, Place, magnitude of the phenomenon Information related to the people for instance, the number of dead, wounded, missing, damaged and affected person. Information related to the infrastructures, the number of destroyed and affected houses Information related to economic impact, such as the number of affected hectares, monetary lost among others

The proposed system architecture uses two main modules: text filtering and fact extraction. The first module concentrates on selecting news report about natural disasters and the second on extraction of information from the texts. Both the modules apply a supervised classification approach, which is a machine learning technique that generates functions from training data that allows mapping input object to a set of predefined categories.

Their approach was tested on Spanish news report about natural disasters. The corpus were collected from several online Mexican news papers from 1996 to 2004 and get an F measure of

98 % in the detection of documents about natural disasters and 76% in extraction of relevant data from those documents.

Most of the related work we review and the theoretical literatures in IE show that the use of knowledge engineering approach by mixing it with the automatic training approach is more efficient than the pure automatic training approach. The simple addition of basic rules to the automatic training approach improves the performance of the IE system highly when compared with pure automatic training approach.

3.6 Information Extraction and Related Works on Amharic Text

Different graduate students of Addis Ababa University and College of Telecommunication and Information Technology (CTIT) have conducted research on NLP in different areas. Among these research works that are more related to our work are reviewed and presented as follows.

3.6.1 Text Categorization from Amharic Text

Among the different categorization researches that are done on Amharic text, the research work presented in [33] uses a machine learning approach to categorize the Amharic news text in to 15 different categories. The Ethiopian News Agency (ENA) news data and the manually given category to the news are used as a corpus. The Weka package that has different machine learning algorithm for classifier learning is employed which are SVM and decision making and has a good result in classifying the news text. The different preprocessing tools are employed on the Amharic text before it is given to the Weka classifier learning algorithms to learn the classifier model. The tools employed are the following

Word Identification: The Amharic word separator which are single space," netela serez", "hulet neteb", "dereb serez", "arat neteb" also written in Amharic as ፣, ፡ , ።, and ፤ respectively carriage return, line feed, tab etc are used to identify words from the Amharic text. The hyphen between words was also deleted to merge them in to a single word like ክፍለ-ከተማ into ክፍለከተማ. The numbers weren't considered in their work as it doesn't differentiate one document from the other.

Stop Word Removal: From the identified words they remove stop words (ነ ው; ኣ ና, and ነ በር) as they doesn't differentiate one document from the other as stop words exist in almost all documents.

Stemming: varieties of words created by language requirement affixes could result in feature words redundancy which in turn reduces the efficiency and accuracy of the text classifier. To solve this problem stemmer is used to removes common suffix and prefix to change the word variant into one common form, which will increase the efficiency of text classifier.

Controlling Spelling Variation: Observation of the Amharic news words that have been spelt in different ways seems to indicate that the cause of the spelling variations is the difference in the Amharic pronunciation of the words. Most of these words are foreign words adapted to Amharic. For example, the word exhibition adapted from the English language has been spelt in the following four different ways in ENA's news items: **ኤግዚቢሽን, ኤግዚብሽን, ኢግዚቢሽን, ኢግዚብሽን.** The spelling variations of words were controlled to have one form of representing the same word since it is represented using different form due to its pronunciation.

Identifying Compound Words: The compound words in Amharic sometimes written as a single word and sometimes as two separate words. This non-uniform way of writing could result in semantic loss during document representation - thereby reducing the accuracy of document representation. They identify the compound words and represent them as a single word instead of two separate words.

Feature Word Selection with Their Frequency: the feature word, its frequency in the documents, the number of documents it exist, weight given to the word is all stored in the database as weka the machine learning package they use for implementation takes ARFF format only. The stored data is then converted in to ARFF format to train the classifier and it is later used to test the classifier model.

Their system is tested on Amharic news text from ENA using the Decision tree and SVM algorithms that exist in Weka package. The experiment is conducted by using decision tree and SVM algorithms among the different algorithms that Weka support. The SVM performs better than the decision tree. The following is the summarized experimental result while using SVM

The average result of the classifier is shown as follows

Classifier	Average accuracy for 5-category data	Average Accuracy for 10-category data	Average Accuracy for 15-category data
LibSVM	95.21%	91.36%	81.15%
LMT	93.45%	89.98%	79.72%

This shows that the SVM, among the different classifier algorithms that Weka support is the best algorithms in correctly predicting the category for the different news texts used as dataset for training and testing the system.

3.6.2 Information Extraction from Amharic Text

Regarding the research on information extraction there is one research work presented [34], which uses the hidden Markov model to extract information from the Amharic text. The IE tool is developed for extracting information from a single sentence. It uses the slots subject, object, action and reporter and tries to extract information from the news text which contain the above listed four slots on a single sentence. If one of these components doesn't exist in the sentence the system doesn't extract the information. The extracted information doesn't have that much usage as the facts with describe the different thing about entities are not part of the research work. In this research work not much emphasis is given in studying the language as it is the important building block of the IE tool.

Summary

The different related works that we reviewed so far uses machine learning or a hybrid approach which is machine learning and knowledge based. It is known in the area of natural language processing researches that Knowledge based approach is good to gain a good performance for different NLP based systems. However, it is very difficult to build knowledge based system as it requires much more time and knowledge expert in the domain of development. It is also not easily modifiable to use it for other languages and domains and to adding new modifications is difficult as it requires knowledge expert. To alleviate these problems different machine learning algorithms are used and better results are acquired that can be compared to knowledge based

system. Text classifier approach to IE also shows a good result in most of the IE researches and it uses off shelf machine learning algorithms which make it easy to develop IE system for language like Amharic which has limited NLP resources. Therefore, the text categorization approach with induced features is used in this work as there are no well developed standard NLP resources for Amharic.

CHAPTER FOUR

4. THE AMHARIC LANGUAGE

This chapter discusses the different issues about the Amharic language that is needed in the development of an IE system and a sub text categorization system for Amharic news text. It begins by introducing the Amharic writing system. The numerals and punctuation marks in Amharic language are also described.

4.1 The Amharic Language

Amharic is a Semitic language spoken in many parts of Ethiopia, a country of 73.92 million people by the as reported in the 2007 census from Ethiopia central statistics agency [45]. It is the official working language of the Federal Democratic Republic of Ethiopia and thus has official status nationwide. It is also 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. Outside Ethiopia, Amhara 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].

4.2 Grammatical Structure of Amharic

Word Categorization in Amharic

The words in Amharic are categorized under five basic categories by Baye Yimam [36] that uses the morphology and position of the word in Amharic sentence as criteria. These five categories are ስም (noun), ግስ (verb), ቅፅል (adjective), ተውሳከ ግስ (Adverb) and መስተዋድድ (preposition) [36].

Noun: a word will be categorized as a noun, if it can be pluralized by adding the suffix ኦች/ ዎች (“owch”) 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 lanaguage.

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

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.

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.

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, ከፉኛ, እንደገና, ጅልኛ and ግምኛ.

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

4.3 Amharic Punctuation Marks and Numerals

In Amharic, there are different punctuation marks used for different purposes [1]. In the old scripture, a colon (two dots ፡) has been used to separate two words. These days the two dots are replaced with whitespace. An end of a statement is marked with four dots (አራት ነጥብ ።) while ነጠላ ሰረዝ (፣ or ፥) is used to separate lists or ideas just like the comma in English. The punctuation marks which are used for sentence demarcation and those which are used for separating similar items are important to this research.

In Amharic, numbers can be represented using either the symbols of Arabic number system or the symbols of the Ethiopic number system or using words and symbols of the Arabic number system. Table 4.1 shows the Arabic, Amharic and alphanumeric representation of numbers.

Table 4.1: Number Representations in Amharic

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

In Amharic, fractions and ordinals have their own way of representation [1]. Table 4.2 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.

Table 4.2: Amharic Fraction and Ordinal Representation

Fraction	Amharic representation	Ordinals	representation
½	ግማሽ	1st	አንደኛ/ቀዳማዊ
1/3	ሲሶ	2nd	ሁለተኛ/ዳግማዊ
¼	ሩብ/እርዖ	3rd	ሦስተኛ/ሳልስ
2/3	ሁለት ሲሶ / ሁለት ሦስተኛ	4th	አራተኛ/ራብዕ
¾	ሶስት-አራተኛ	.	.
1/10	አስራት	.	.
2X	እጥፍ	9th	ዘጠነኛ/ዘጠነ ኛ
2.X	ሁለት ነ ጥብ	10th	አስረኛ

Dates in Amharic can be written in different ways. It can be written using symbols in Arabic number system like 12/01/2001 or using Ethiopic numeral representation and alphanumeric representations like ጥቅምት 10/2003. Knowing the Amharic fraction and its representation in Amharic language is important as they are mostly used in the Amharic text in the form of words instead of digits.

4.4 Characteristics of Amharic writing

The characteristics of the Amharic writing system considered in this section are limited to those that are common to news texts and IE systems.

Character Redundancy: Amharic took the whole Geez alphabet and uses it in the Amharic writing system. It then added some more symbols for some other sounds that it has and that could not be represented by the symbols of the Geez alphabet. This unsystematic borrowing from Geez has resulted in redundant characters in the Amharic FIDEL [33]. Table 4.3 shows an example of the character redundancy where more than one symbol is used for same sound.

Table 4.3 Amharic Characters with Same Sound

Consonants	Other symbols with the same sound
ሀ (hä)	ሃ ሐ ሓ ኃ ኀ ኻ
ሰ (sä)	ሠ
አ (ä)	ኣ ኡ and ዓ
ጸ (tsä)	ፀ

Spelling variations of a word would unnecessarily increase the number of words representing a document which could reduce the efficiency and accuracy of the classifiers for sub text categorization as well as for the classifier used for the IE system. Amharic document processing for feature selection should therefore normalize word variants (spelling differences) caused by inconsistent usage of redundant characters. During the pre-processing stage of Amharic documents in this work, the different forms of a character that have the same sound are changed to one common form. Table 4.4 shows examples of the different word spellings caused by the redundant characters.

Table 4.4 Word Spelling Variations

The Word in English	The word in Amharic	Spelling Variants of the Word
Work	ሥራ	ስራ
Sun	ፀሐይ	ጸሐይ፣ ፀሀይ፣ ጸሀይ
World	አለም	አለም፣ ዓለም፣ ኣለም
Power	ሀይል	ሃይል፣ ኃይል

Compound Words in Amharic Language

In the Amharic writing system, inconsistency is often observed regarding the representation of compound words. Some compound words are used as a single word in some instances (either by fusing the two words or by inserting a hyphen between them) and as two separate words at other instances. Inconsistent usage of compound words could result in redundant word features by

creating more words. when a compound word (example አዲስ አበባ) is treated as two separate words አዲስ and አበባ.

Variations due to Pronunciations: usage of foreign language words in Amharic is also found to be another source of word spelling variations. Observations of the ENA's news documents shows that in most cases in the writings of words adapted from foreign languages different writers use different spellings. 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. For example, the word ሚቲዎሮሎጂ (Meteorology) have different Amharic spellings. Table 4.5 shows examples of spelling variation in the writing of foreign words in Amharic.

Table 4.5 Word Variations due to Pronunciation

Foreign Word	Equivalent Words in Amharic usage
Meteorology	ሜትሪዮሎጂ፣ ሜትዎሮሎጂ፣ ሜትዮሮሎጂ፣ ሚቲዎሮሎጂ፣ ሚቲዎሮሎጅ፣ ሚቲዎሮሊጂ፣ ሜትዎሮሎጂ፣ ሜትሮዎሎጂ፣ ሜትሮዎሎጅ፣ ሜትሪዎሎጂ፣ ሜትሮሎጂ፣ ሜትር ዎሎጂ፣ ሜቲዎሮሎጂ፣ ሜቴዎሮሎጂ
Million	ሚሊዮን፣ ሜልዮን፣ ሚሊየ ን
Television	ቴሌቪዝን፣ ቴሌቪሾን፣ ቴሌቪዝን፣ ቴሌቪዝን

Moreover there are word spelling variations that could be attributed to variations in pronunciations at different parts of the country, like for example using the two words ጠባይ and ፀባይ to mean temperament or using the three words ጠንዚዛ, ጠንዝዛ and ጥንዚዛ to mean beetle [33].

Other Cases of Word Variations: Difference in word affixing has also been observed to cause word spelling variations. For example difference in suffixing would result in the two writings አእምሮአዊ and አእምሯዊ to refer to human intellect while difference in prefixing would give the two writings ለአንድ and ላንድ to mean 'for one'.

4.5 Sentences in Amharic

A sentence, in every language, is a group(s) of word(s) that comply with the grammatical arrangement of the language and capable of conveying meaningful message to the audience. 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.

Declarative Sentence: is a sentence that is used to express the physical, psychological, imaginary or real events. It is main objective is description of some issue. Most of the sentences are declarative sentences. The news articles use the declarative sentence for expressing different information on different issues. There are also Interrogative Sentence which is used to ask a question, Exclamatory Sentence which is used for emphasis and emotion, and others.

Summary

The Amharic news text which we consider for training and testing of our system uses declarative sentences to describe different issues about infrastructure news. Most of the facts that we try to extract in this research work are names which are name of a person, name of an organization, name of a place, numbers which describe different facts and date and times in the text. These different facts in the Amharic text are either a noun or number. The ግስ (verb), ቅፅል (adjective), ተውሳክ ግስ (Adverb) and መስተዋድድ (preposition) are not extracted in IE as they are descriptive words to the noun and what the noun does or what is done on the noun.

CHAPTER FIVE

5. THE AMHARIC TEXT INFORMATION EXTRACTION MODEL

In this chapter we will present our proposed Amharic Text Information Extraction (ATIE) Model (Figure 5.1). The main components of the ATIE model along with their subcomponents and the interaction between the main components will be presented.

5.1 Components of ATIE

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 [2]. In addition to these three components other subcomponents are also included in each of the main components. The Amharic Text IE model (ATIE) designed in this work has one additional component called Text Categorization in addition to the above listed three components. These four main components of ATIE also contain different subcomponents which are language specific and general subcomponents that are required in IE. The main components interact with each other to make information extraction from Amharic news text possible.

The document preprocessing component is responsible for normalizing the different language specific features which are caused due to its writing system. The language specific issues like normalization of characters, normalization of numbers and tokenization are done at the document preprocessing component. The text categorization component on the other hand handles the categorization of preprocessed news text in to predefined categories. The categories for the news which will be predicted using the text categorization component are investment, infrastructure and others, which are among the subcategories under economy news main category. As IE is a domain specific task text categorization component is used to select the specific news text that can be used as an input to the IE system. The learning and extraction component extracts candidate texts from the news text, which will be used as attributes to fill the database slot and learns a classification model that will be used to predict the category of the candidate text. The post processing component is responsible for the formatting of the extracted data according to the structure of the predefined database slots.

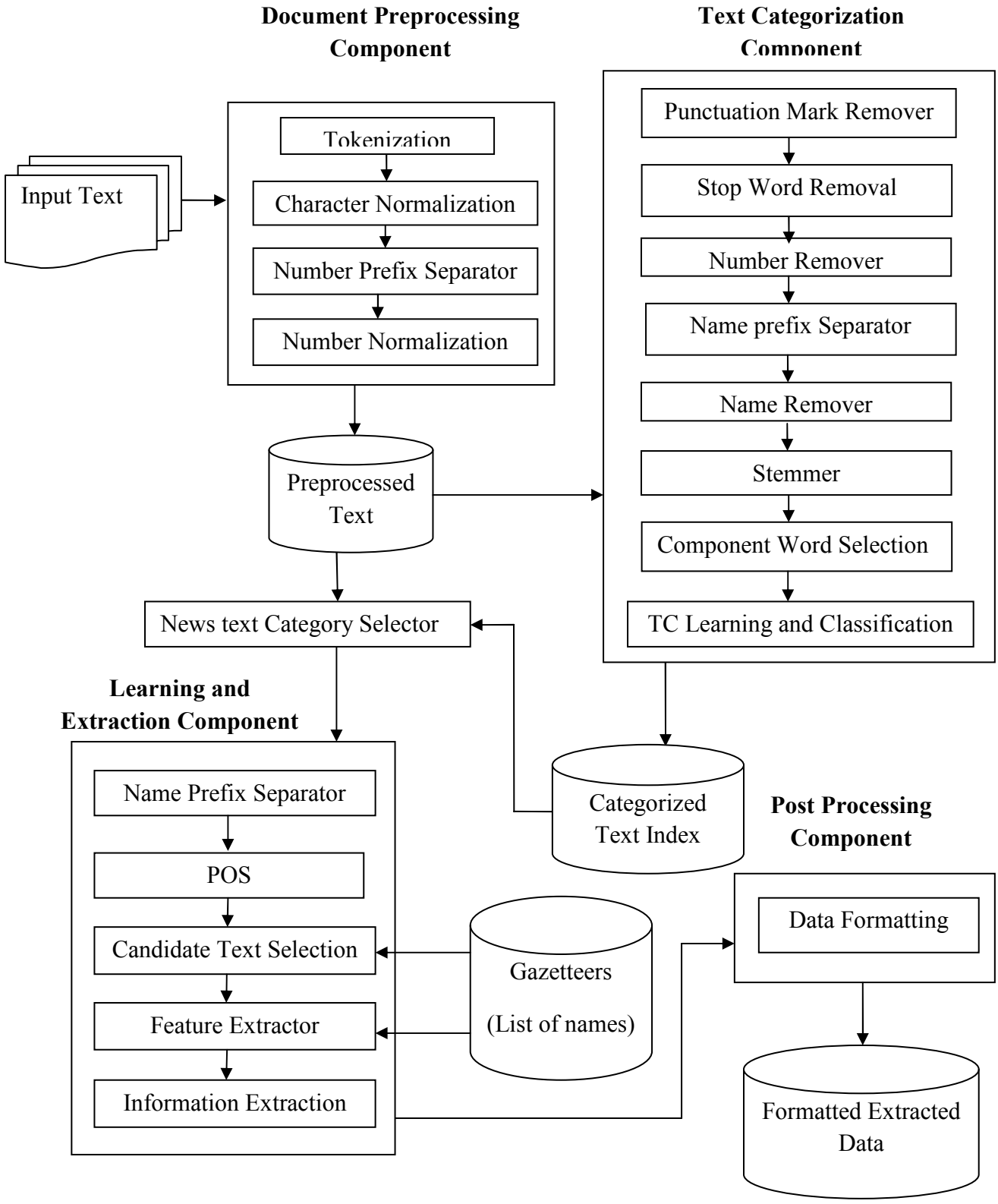

Figure 5.1 The proposed Amharic Text IE Model

In the following subsections we give a more detailed description of the four main components along with their subcomponents of our proposed ATIE model.

5.2 Document Preprocessing

The documents used as a corpus, training data and testing data for this work are Amharic news texts obtained from ENA. The Amharic news texts need different type of preprocessing before they are made ready for ATIE system as there are different language specific writing that should be normalized. The document preprocessing component consists of tokenization, character normalization, Number Prefix Separator, and number normalization as subcomponents to handle language specific issues. These subcomponents and their function are described as follows.

Tokenization: is the task of chopping up the given text into pieces, called tokens [1], which are then used by text categorization and IE. Since the text categorization components uses a collection of words called features (attributes) to represent the news texts and IE uses token and their POS as features for training and testing of the classifiers, tokenization is crucial for both components. The tokenization is then used to chop up the news text into meaningful tokens that can be used as features for both text categorization and IE. The Amharic punctuation marks, the space between words are used for the tokenization process.

Character Normalization: The Amharic language has different characters with the same meaning and pronunciation but with different symbols. The letters such as ሠ and ሰ ; ሀ, ኀ, ሃ, ኻ, ሐ, ኃ and ሐ and ጸ and ፀ are examples of characters with the same meaning and pronunciation but different symbol. Using them as they appear in Amharic literatures might have a meaning from the language point of view but they need to be normalized when developing NLP tool for Amharic language. These characters should be normalized to a single characters like ሠ and ሰ to ሰ and ሀ, ኀ, and ሐ to ሀ and ጸ and ፀ to ፀ as well as their orders (ሠ, ሡ, ሢ, etc. to ሰ, ሱ, ሲ, etc.) accordingly. In addition to this normalization, 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 appear in the news texts will create different words that will be considered as different from one another in the training and testing of the classifier for text categorization and IE. The role of character normalization is to normalize all these characters of the same sound and meaning but different shapes to a single character.

Number Prefix Separator: Numbers in the Amharic news text appear using the number and word representation. For example, the number might appear as ከ2ሚሊዮን, ከ2 ሚሊዮን, የ2ሚሊዮን, የ2 ሚሊዮን, በ2ሚሊዮን, and በ2 ሚሊዮን. This kind of number representation will cause a problem during number normalization as they cause confusion in identifying them as numbers. The role of Number Prefix Separator is then to separate the characters ከ, በ, and የ, which are the mostly attached prefixes from the number, which will make the news text ready for number normalization.

Number Normalization: the numbers in Amharic text appear by using the most widely used number system in the world which is the Arabic number system. The problem is the numbers in the Amharic text are mostly represented using words and digits. For example in most of the ENA news the number 2,500,000 is written as "2 ነጥብ 5 ሚሊዮን" which is equivalent with 2.5 million. The following are among the possible number representations in Amharic news texts 1ሚሊዮን, 1ሺ, 1ሺህ , 1ቢሊዮን , 1ሺህ 150, 1ሺህ 40, 1ሺህ 4, 1 ሺህ 1 መቶ ሃ ምሳ , 1 ሚሊዮን 1 መቶ ሺህ , 1 ቢሊዮን 5 ሚሊዮን 400 ሺህ , አንድ ቢሊዮን አምስት ሚሊዮን አራት መቶ ሺህ , 1 ነ ጥብ 5 ሚሊዮን , and 1 ነ ጥብ 6 ቢሊዮን . These and other possible number representation in the news text causes problem during text categorization and IE as it is difficult to uniquely identify the numbers in the news text. The number normalizer then changes all above listed of number representation in to their equivalent number representation. For example "2 ነጥብ 5 ሚሊዮን" will be normalized in to 2500000.

5.3 The Text Categorization Component

The text categorization component manages to categorize the news text as one of the predefined categories by using a trained classifier model. It categorizes news texts so that the infrastructure category news text will only be used as an input for IE. There are different subcomponents under the text categorization which are used to preprocess the news text for the selection of words that will be used as features (attributes) for training and testing of the classifier. The different subcomponents of text categorization are punctuation mark remover, stop word removal, number remover, Name Prefix Separator, name remover, stemmer, component word selection and text categorization.

Punctuation mark remover: punctuation marks which exist in the news texts are not important for text categorization. Therefore, the different punctuation marks that exist as independent token or attached to another token are identified and removed.

Stop Words Removal: the words like አስታውቀዋል, አስታውሰዋል, እስካሁን, አሳሰበ, አሳስበዋል, አስፈላጊ, አስገነዘቡ, አስገንዝበዋል, አብራርተዋል, አስረድተዋል and others listed in Appendix A, which are called stop words are not important for text categorization. Since they exist in every category of Amharic news texts they are not important to uniquely identify the one category of the news from the other. Using them for text categorization will degrade the performance of the classifier. Therefore the stop word removal removes all the above listed and other list of words in appendix A, which are considered as stop words in the Amharic language from the news texts.

Stemming: reduces words to their root word so that the root word will be used for text categorization. Ignoring removal of suffixes and prefixes from the word will increase the number of words used as a feature for categorization which will consecutively degrade the performance of the classifier. The purpose of stemming is to reduce all words to their root words so that the classifier will be efficient since it uses the root words as features.

Number Remover: the different numbers which represent different aspects in the news texts or any other text can't uniquely identify a text as they can appear in different news and can be used in different ways. The purpose of number remover is to remove all the numbers that exist in the given text so that the number of features used to represent the news text will be minimized and unique. Any token which contain numbers and any token which is a number will be removed by using the number remover.

Name Remover: The names of a person, place, organization or any other type of names exist in different news represent a different role than they have on other news. For example, the person name in investment news might be used as the politician name in politics news or a place listed in infrastructure news might also exist in sport news for conducting a national football match. Therefore the names might be used in different news having different roles which makes them not a feature to uniquely identify one news text from the other. The name remover then removes all the names of a person, place, organization and titles attached to person name. The place and organization names which exist in the news text mostly have prefixes attached to them which

make it difficult to remove names for that we developed another preprocessing tool called Name Prefix Separator.

Name Prefix Separator

Most of the place and organization names which appear in the different ENA news we have used as a corpus for the prototype development of our system has prefixes like ዐ, የ, ከ, and ኦ attached to them. These prefixes are identified after the analysis of the news texts considered as a corpus. The function of the Name Prefix Separator is then to stem the prefix of all the names if they have one of the above listed prefixes.

Component word selection: after the Amharic news text is preprocessed, the stop words, names and numbers are removed and the remaining words are then used as features for text categorization.

Text Categorization Model Learning and Classification: this subcomponent deals with training a classifier model and uses it for prediction of the category of the unseen news texts. The component words are used as an input for this subcomponent. The Weka preprocessing tools and different classifiers which exist in Weka are used to train the classifier model. The selected component words are then converted to ARFF format which is the dataset format that Weka supports.

5.4 Learning and Extraction Component

This component is the main part of the ATIE model which is responsible for extracting candidate texts from the infrastructure news texts and for learning a classifier model to predict the category of the candidate texts, which will be used as an attribute for the database slots. This component uses both the output of text categorization component and document preprocessing component. The news texts which are classified as infrastructure news only passed to the learning and extraction component and the other news are discarded as they are not considered for extraction due to the domain specific nature of IE. The extraction and learning component also comprises of different subcomponents which will be used to make the data ready for extraction.

Part of Speech Tagger: assigns a POS tag to each token. The set of tags includes the conventional part of speech in Amharic such as noun, verb, adjective, adverb, prepositions and pronoun. It is important for extraction as it is one of the features considered.

Candidate Text Selection: identifies possible candidate texts that are going to be extracted from the news texts. The names of a person, place, organization, infrastructure and the different numbers which exist in the news text are considered as candidate texts for extraction in this research work. The candidate text selection then selects all the above listed text segments from the news text as a candidate attribute values for extraction and tag them according to their category that is as a name or a number.

Feature Extractor: extracts different features to uniquely identify the candidate text as one of the predefined attribute. The features used for IE are POS tags, immediate attributes on both side of the candidate token, the prefix and immediate next token of candidate token, the token category of the candidate token which is the name of a person, place, organization, and infrastructure or a number are extracted by the feature extractor. The extracted features by feature extractor will be stored in the database for training and testing a classifier model.

Information Extraction: Any NLP tool developed for any language if it uses the machine learning approach will have training and testing phase. The information extraction subcomponent also has the training and testing phase. The following figure shows the detailed presentation of Information Extraction subcomponent

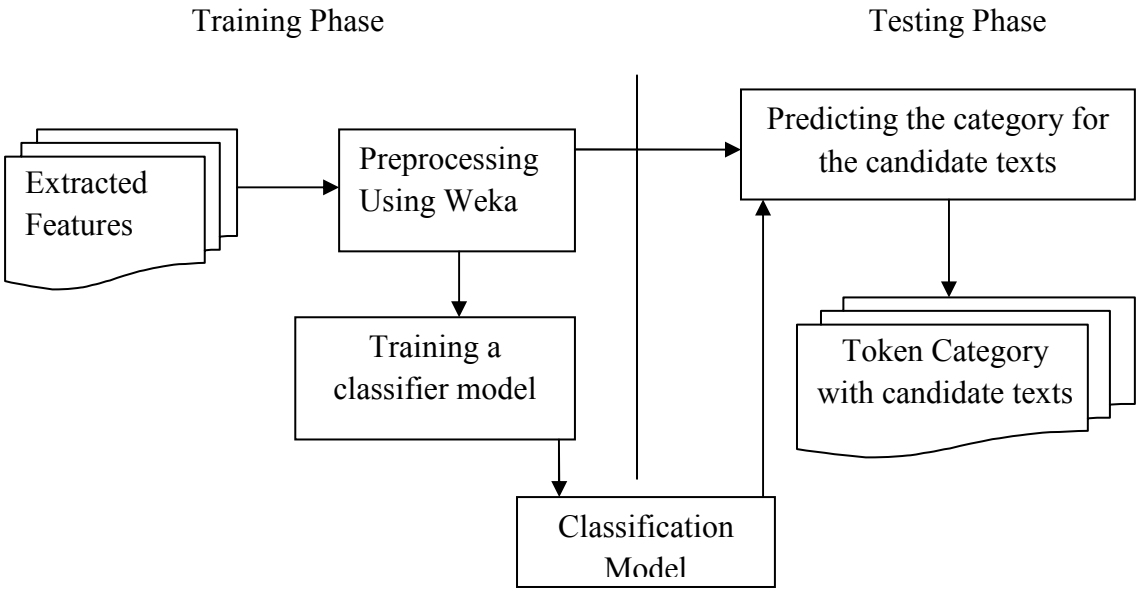

Figure 5.2 The Information Extraction Subcomponent

During training phase the classifier model is generated based on the extracted features from tagged training data. The extracted features are then preprocessed using Weka to make the data ready for training a classifier model. After preprocessing, the classifier model will be generated which will be used for predicting the category of the candidate texts during the testing phase. The output of the information extraction subcomponent is the candidate texts with their respective predicted category, which is used as an input to the post processing component.

5.5 Post Processing

It is the last component of ATIE model which formats the different attributes that are extracted from the news text and store them in the database according to the predefined format of the data base slots. The main function of the post processing component is to arrange the format of the extracted data so that it will be flexible for data mining or any other application which want to use the data.

Summary

In this chapter the ATIE model is presented and the main tasks of the different components are described. The ATIE model comprised of document preprocessing, learning and extraction and post processing and an additional text categorization component, which categorizes the Amharic

news text as one of the predefined categories. The document preprocessing handles the processing of language related issues, the learning and extraction component learn a classifier model and extract the candidate text segments from the news text, and the last component of the system handles the formatting of the data to store it in the database. The ATIE model that is proposed in this research work is a generic model which can be used for any other domain in the Amharic language as long as there is a sufficient annotated training data.

CHAPTER 6

6. IMPLEMENTATION

In this chapter, an Amharic text IE prototype system that is developed based on the proposed model will be presented. The resources and algorithms used in the four main components of the system and how the candidate texts for extraction are identified from unseen news texts are also discussed.

6.1 Data Source

The data that we use to train and test ATIE system is obtained from Ethiopian News Agency (ENA). ENA has a database containing different Amharic news from year 2005 to present. These Amharic news texts are categorized under 16 main categories manually which are law and justice, health, events directory, international relations, social affairs, culture, politics, agriculture, defense and security, science and technology, sport, education, economy, accident, weather and other classes. The main reason for selecting ENA as a data source is the availability of large collection of manually categorized news texts in to one of the sixteen categories which make it easy to select the specific news from the available category for training and testing purpose.

Among the different categories from ENA news the economy category is used as a data source for the training and testing of our ATIE prototype system. The reason for selecting economy news category, among the other categories is the availability of factual information which can be extracted and stored in the database. During selection of the news category as a data source for our system we consider all the sixteen categories and analyzed which one of these categories contain facts that can be extracted and stored in the database which can be used by different application and individuals after extraction. Most of the other news categories are more subjective which have less factual data that need to be extracted. The other reason is the importance of the data after extraction. Economy news contain many vital information that investors, financial organization or any individuals can use and extracting these information and using them for later stage or by other application will be vital. Due to these reasons and the

nature of IE which is very domain specific the Economy main category and infrastructure subcategory news texts are used as training and testing dataset in this research work.

6.2 Document Preprocessing

The document preprocessing component handles the different language specific issues that are imposed by the nature of the language to make the data ready for remaining phases. The data collected from ENA requires different normalization before it is used by the ATIE system. There are different reasons for including document preprocessing in our system. One of the reasons is the nature of the Amharic language characters which sound the same and used for the same purpose but with different symbols. Using these characters as they appeared in a text without normalization will make redundancy of words which in turn degrade the performance of text categorization and IE as words are the features used for categorization purpose in both components. The other reason is that number representation in the Amharic news text is complicated. In some of the news it is represented using digits and in other news it is represented using digits and Amharic characters and in some other news it is represented using characters only. If we use them as they are they will create confusion during text categorization and IE as it is difficult to consider them as numbers or words. The different document preprocessing components are described as follows.

6.2.1 Tokenization

The purpose of tokenization in this research work is to break down a stream of text into words, phrases, and symbols which will be used as meaningful elements called tokens for text categorization and IE. As tokenization is a preprocessing subcomponent which will be used by both text categorization and IE component, it is developed in a way it satisfy the requirements of both components. The tokenization uses the ። (አራት ነጥብ) the Amharic full stop, and ፣ (ነጠላ ሰረዝ) the Amharic comma as the most commonly used punctuation mark in the news texts. These punctuation marks play a vital role for extraction. The features for the tagged text are extracted at the sentence level. The ። is used for identifying the sentence demarcation and ፣ is used to separate different text segments which mostly are related. The tokenization then tokenizes all the text segments which have space between each other as independent token. ። is used as

independent token and ፣ separated words are considered as a single token as they represent a list of things under the same category.

6.2.2 Character Normalization

In Amharic languages there are characters which have the same sound and can be used interchangeably but have a different symbols. Using these characters as they are, might have a meaning. But when we come to implementing the Amharic text in natural language processing, using them as they appear will reduce the efficiency of the system. For example, if we use them as they are for text categorization we will have the same word with different representation as a feature word which will degrade the performance of the system. If these characters are used as they appear without normalization the same sound words will be written using different symbol characters for example ጸሀይ፣ ፀሀይ፣ ጸሃይ፣ ፀሃይ which represent the same thing "sun" but appears in different format will be used as independent features. Normalization of these characters in to a single character is task of character normalization. The following are a list of characters with the same sound but different shapes.

Table 6.1 Amharic Characters which have different symbols but similar sound

Characters with the same sound	Normalized to
ሀ, ኅ, ሐ, ኃ, ኻ, ሓ, ሃ	ሀ
ሰ, ሠ	ሰ
ጸ, ፀ	ጸ
ቸ, ቼ	ቸ
ች, ቺ	ች
ሸ, ሼ	ሸ
የ, ዬ	የ
ዩ, ይ	ይ
አ, ኣ, ዐ, ዓ	አ
ኝ, ኚ	ኝ
ሽ, ሺ	ሽ

6.2.3 Number Prefix Separator

Numbers in the infrastructure news text are one of the facts that are considered for extraction in this research work. The value for the attribute number of users and amount of money spent for the development of the infrastructure are numbers. According to our analysis of the different infrastructure news numbers have the two representation either they are attached to prefixes or exist as independent tokens. The purpose of Number Prefix Separator is to stem the prefix from the number and to consider the prefix and the number as independent tokens.

6.2.4 Number Normalization

Numbers are one of the text segments which are extracted by the ATIE system as facts to be stored in the database. As the economy news texts are used as a data source for this research work the different numbers representations in these news texts are considered. Numbers in these news texts are represented using digits and words. For example the number 1.5 billion will exist in the news text as "1 ነጥብ 5 ቢሊዮን" is needed to be extracted during IE and removed during text categorization as it is not needed. Normalization of these number representations will make the task of extracting the numbers for IE and removing for text categorization easy. The function of number normalization is then to normalize all the numbers in the Amharic news texts which are represented using digit or Amharic language characters or a combination of both in to their equivalent number representation. The number "1 ነጥብ 5 ቢሊዮን" will be represented as 1500000000 after normalization. The following is the algorithm for number normalization.

For each news document in the corpus

Check for the presence of number or a word representation of a number or combination of words and numbers existing as a single token

If present and the word is attached to a number

- Split the word or character from the number
- *Initialnumber* = the number separated from the character or word
- *Candidatetext* = the number representation of the separated word or character or if it is not a number the character itself
- *Token1*= the immediate next token
- *Token2*= token next to token1

- *Token3*= token next to token2
- *Token4*= token next to token3
- *Token5*=token next to token4

Else if the word and the number are independent tokens

- *Initialnumbe*r = the number presented
- *Token1*= the immediate next token
- *Token2*= token next to token1
- *Token3*= token next to token2
- *Token4*= token next to token3
- Token5=token next to token4

End if

If Candidatetext is a word representation of a number

Candidate text will be the number equivalent representation of the word segment

Else if Token1 is a word equivalent representation of a number

Token1 will be the number representation of the word

Else if Token2 is a word equivalent representation of a number

Token2 will be the number representation of the word

Else if Token3 is a word equivalent representation of a numbr

Token3 will be the number representation of the word

Else if Token4 is a word equivalent representation of a number

Token4 will be the number representation of the word

Else if Token5 is a word equivalent representation of a number

Token5 will be the number representation of the word

End if

If Candidatetext is "NETB" or "." and Token1 is a number and Token2 is a number and Token3 is not a number

The normalized number will be the merger of Initialnumber, Token1 and Token2 by removing one zero from Token2

Else If Token1 is not a number and Newnumber is not null

The normalized number will be the Merge InitialNumber and Newnumber

Else if Token1 is not a number and Candidatetext is null

The normalized number will be initial number

Else if Token1 is a number and Toekn2 is not a number

If the number of digits of Token1 is two

The normalized number will be Initialnumber+0+Token1

Else if the number of digits is one

The normalized number will be Initialnumber+00+Token1

Else if the number of digits are three

The normalized number will be Initialnumber+Token1

End if

Else if Token1 is a number and Toekn2 is a number and Token3 is not a number

The normalized number will be the Initialnumber+Token1+Token2

Else if Token1 is a number and Token2 is a number and Token3 is a number and token4 is not a number

The normalized number will be the merger of Initialnumber, Token1 and Token3

Else if Token1 is a number and Token2 is a number and Token3 is a number and Token4 is a number and Token5 is a number

If number of digits of Token3 are three

The normalized weight will be the merger of Initialnumber, Token1, Token3, and Token4

Else if number of digits of Token3 are two

The normalized weight will be the merger of Initialnumber, Token1, the digit 0, Token3 and Token4

Else if the number of digits of Token3 is one

The normalized weight will be the merger of Initialnumber, Token1, the digit 00, Token3 and Token4

End if

Else if

End if

End for

Figure 6.1 Algorithm for Number Normalization

6.3 Text Categorization

The text categorization component of ATIE prototype system is used to categorize the news text as investment, infrastructure or others categories. The need of categorizing news texts is because of the domain specific natures of IE. The different processes that are involved in the text categorization component are organized under the following three activities

- Training data preparation
- Training a classifier model
- Using the trained classifier model for text categorization

6.3.1 Training Data Preparation

The training data that is used for text categorization is economy news category which is obtained from ENA. The different news texts which are under the investment, infrastructure and other sub categories are used for training and testing the text categorization. The training data is prepared by manually selecting all the news texts and storing them in the text file format. Then, all the text file format news texts are organized in three different folders which have the same name to the category of the news texts which are investment, infrastructure, and others. To obtain features which will be used for training and testing the classifier the following preprocessing tools are

used for each of the news texts. For the text categorization subcomponent the work of Yohannes Afework [33] which is done on text categorization using Weka is adopted.

Feature Word Selection

For the purpose of text categorization the news texts are considered as a bag of key words which are called features or attributes of the document [3]. Not all words in a document are considered as features. The purpose of feature selection is then to reduce the dimension of the data by selecting features from the original words of the news text. Different preprocessing tools can be used on the news text for feature selection. Redundant attributes are identified and removed as a result of feature word selection process. The following processes are used for facilitating the feature selection from the news text.

Stop Word Removal

Words that can't represent the documents since they commonly exist in different news texts and sentences are considered as stop words. If a word exists across documents it is considered as a stop word in text categorization. Words like አስታውቀዋል, አስታውሰዋል, እስካሁን, አሳሰበ, አሳስበዋል, አስፈላጊ, አስገነዘቡ, አስገንዝበዋል, አብራርተዋል, አበራርተው, አስረድተዋል are called stop words. This and other stop words which are listed in the research work of Melese [40] are considered as stop words and removed as they are not important for text categorization.

Stemmer

Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root word form [1]. Stemming is an important part of text categorization for language like Amharic which has many morphemes for the same root word. Words might be expanded to represent number, feminine, ownership, and other issues but for text categorization they are all the same. Using all the words as they exist will degrade the performance of text categorization as the feature words used for classification increases. For example, the words ልማትና, ልማትንና, ልማትንናም are changed to their base term ልማት '*lemat'* (development). Similarly the variants of selam such as ሰላምና, ሰላምንና, ሰላምንም are changed to ሰላም *'selam'* (Peace). The stemming algorithm developed in [41] and later used by Seid [1] is adopted for our work as it is developed in the way that suits our work.

Number Remover

Numbers which exist in different Amharic news representing different amount associated with them are not important for text categorization. The numbers exist in economy news might also exist in sport, social or any other news category. Numbers are then not used for text categorization. Number remover removes all the numbers and tokens which contain numbers from the given news text.

Name Remover

In Amharic news text names of a person, place, organization, titles of a person exists having a different meaning from one document to the other. The person names which exist in the economy news as an investor might exist in other news having different role. Using names as a feature for text categorization is not important as it doesn't uniquely identify the news document. The name remover then removes all the names by using the Gazetteer lists. The problem in removing the names from the Amharic news text is that these names might have prefixes. Most of the place and organization names in the news text that we use have one of the following four prefixes ከ , ወ, ለ, and የ attached to them. To solve this problem the Name Prefix Separator is applied before the number remover is used, which removes all the names in the news text.

Name Prefix Separator

The Name Prefix Separators stem all the prefixes from names. If the first letter of the token is ከ, ወ, ለ, or የ the remaining part of the token is checked if it is a name or not. If it is a name it will be Separated and both the prefix and the name will be removed.

Preprocessing using Weka

After the news texts are preprocessed using the language specific developed tools the next step is to preprocess it using Weka to make the data ready for training the classifier and using the model for prediction at later stage. Since the feature words that are used as dataset for training and evaluating attributes are Strings the *TexttoARFFLoader* tool which Weka support is used to convert all the text files news texts in to ARFF format. A string attribute can have in principle infinite number of values and therefore it cannot be handled by any classifier in Weka. That is why we have to convert string values into a set of attributes that represent the frequency of each

word in the strings. The *StringToWordVector* filtering tools also performs *TF/IDF* transformation. *TF/IDF* weight (Term Frequency–Inverse Document Frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. The *StringToWordVector* filter is then is used to select the feature words that can be used as attributes for training the text classifier [42].

6.3.2 Training a Classifier Model

Once the news text is preprocessed and the features are selected the next step is to train the classifier model which later will be used to predict the category of unseen news texts. For training a classifier Weka open source machine learning algorithm is used. Weka is an open source machine learning algorithm for data mining task. It has open source machine learning algorithm which contains tools for data pre-processing, classification, regression, clustering, generate association rules, and visualization. All of the packages are developed in java at the University of WAIKATO in New Zealand [42].

To use the Weka package one of the three interfaces that Weka provides can be used

- A command line interface, which accept a command through the simple CLI command line interface. There are different commands that can be used with simple CLI.
- An Explorer GUI interface, which has the menus that the user can select for preprocessing, classification and regression and apply them on the appropriate dataset.
- An experimenter GUI interface: which allow running different algorithms in batch and comparing the results

Weka provides different machine algorithms which can be applied on textual data and numeric data. Among these different algorithms Naïve Bayes, SMO and Decision tree are used as they are the most commonly known and used algorithms. The preprocessed dataset is then used to train and evaluate the three classifiers that are used in this research work.

6.3.3 Using the Trained Classifier Model

After the classifier model is generated it can be used for predicting the category of the unseen news text. To use the model to predict the category of unseen news text, the news text should be preprocessed first in a same way the training data is prepared and converted in to *ARFF* format. After that it will be loaded to the Weka system and the saved classifier model is then used to predict the news category for the unseen news text.

6.4 Learning and Extraction

The learning and extraction component main task is to extract candidate texts and train and use the classifier model for predicting the category of the extracted candidate texts. The main hypothesis of this research work is that we human beings identify facts from a document by using the informative words that come before and after the candidate text which we consider are facts. For example, if we see the number in the text to know what it represents, we don't need to read all the document just looking at the words that are on both side of the number will be enough. The main task in the learning and extraction component is then to extract candidate tokens in the infrastructure news text and use the tokens that are on both sides of the candidate token, their respective POS tags, and a token category as features to know the category of the token. The classifier model can use the features to learn how to classify the candidate text as one of the predefined attributes for extraction.

To train and use the classifier model for the learning and extraction component use the training and testing phase. The training phase uses the manually annotated news data to train a classifier model which will be used for extraction while the testing phase evaluates the performance of the trained model on the unseen news texts.

6.4.1 Training Data Preparation

One of the time taking and tedious task of developing an IE system is preparation of training data. For the sake of developing the prototype and experimenting the ATIE system the infrastructure news subcategory which is under the economy news main category is used. But the architecture that we develop for the ATIE system is generic one which can work on any domain of the Amharic news text or any other text as long as there is plenty of a manually annotated data is available for training and testing the system.

The predefined attributes that we set to extract from infrastructure news are the Infrastructure name, Place where the infrastructure is built, the amount of money used to build the infrastructure, the source of money for the infrastructure development, the number of users which will be benefited from the infrastructure, the person who give the information to the news agency. These six attributes are selected as an attribute after analysis of the different infrastructure news and the common facts which exist in most of the infrastructure news and which we thought are relevant facts that should be extracted.

During training data preparation all the text segments which are numbers and names are tagged accordingly. If the name is the fact to be extracted it will be tagged according to the tag set for the predefined attribute otherwise it will be tagged by other tags. The main purpose of tagging all the text which are not facts considered for extraction to train the classifier to identify texts that are part of the predefined database slot and those that are not.

- B-INF for infrastructure name
- B-PLACE for the place
- O-PLACE for place which exist in the text but are not the facts considered for extraction
- B-FUND for the amount of money spent for the development of the infrastructure
- O-NUM for numbers which are not tagged as fund or number of users
- B-FSOURCE- for the financial source for the infrastructure development
- O-ORG organizations which are not the financial source
- B-NUSERS- for the number of users after the infrastructure development is finished
- B-REPORTER- for the name of the person who give the information to the news Agency
- O-REPO- person names other than the one who give the information

6.4.2 Processing of the News Texts for Feature Extraction

To make the news texts ready for IE different preprocessing are required. The features that are used for IE are candidate token, prefix of the token, previous token, immediate next token, next token, POS of the above listed tokens and the token category of the candidate token. The value

of the token category is a number, place name, person name, or organization name as the candidate texts considered for extraction are these. To extract these features from the news texts different processing tools are developed.

POS

The part of speech tagger is one of the features considered for IE. It is developed by using the standard Amharic corpus from Walta. The Hidden Markov model is used to train the POS from the tagged Amharic news text. The performance of the classifier is also evaluated using Ling pipe experimentation option and it assign for the correct POS 79% tokens used for testing.

Name Prefix Separator

The Name Prefix Separator is used to stem the prefix from names. As place names, person names, and organization names are the candidate tokens for extraction the prefix attached to these names is stemmed and stored independently. Person names doesn't have any prefix attached to them which need stemming while organization and place names have a prefix attached to them. Based on our analysis from different corpus the prefixes that are attached to place and organization names are the letters ከ, ባ, ለ or የ. The role of Name Prefix Separator is then to stem the prefix from the name and write them to the text file as independent tokens. The following is an algorithm for stemming prefix from names

For each token in the corpus

Check if the token is a name by using the gazetteer list

If the token is a name

Write the token to a file

Else if the token is not a name

Check for the first character of the token

If the first character is ಹ, ಂ, ಽ or ೆ

Stem ಹ, ಂ, ಽ or ೆ from the token

Check if the remaining part of the token is a name

If true

Write the prefix as independent token

Write the remaining part of the token as independent token

Else

Write the token to the text file as it is

End if

End if

Else

Continue to the next token

End if

End for

Fig 6.2 Algorithm for Name Prefix Separator

6.4.3 Feature Extractor

The feature extractor component is developed to extract the different features and a category from the tagged training data. The features that are used for ATIE system for prediction of the category of the candidate texts are:

- POS of the candidate token

- POS of the prefix
- POS of the previous token
- POS of the immediate next token
- POS of the next token
- Candidate token category
- Prefix token
- Previous token
- Immediate next token
- Next token

The feature extractor extracts all these features from the tagged training data and stores them in the database for latter processing by the Weka preprocessing component and classifier learning. These features are selected after analyzing the nature of the language and by considering it will provide a good prediction for the category of candidate tokens. The features are extracted on a sentence level basis. The following is an algorithm developed for feature extraction in this research work.

For each token in the news document

Check for the presence of a token which is tagged

If present

If the token is the first token

Previous token will be null and its POS will be null

Prefix will be null and its POS will be null

Immediate token will be the first next token to the tagged token and its POS will be the POS attached to it

Next token will be the token next to immediate token and its POS will be the POS attached to it

Else if the token is the second token

Previous token will be null and its POS will be null

Prefix will be the first token and its POS will be the POS attached to it

Immediate token will be the first next token to the tagged token and its POS will be the POS attached to it

Next token will be the token next to immediate token and its POS will be the POS attached to it

Else if the next token is :: (the Amharic full stop)

Previous token will be null and its POS will be the POS attached to it

Prefix will be the first token and its POS will be the POS attached to it

The immediate next token will be :: and its POS will be the POS of ::

The next token will be null

Else

Previous token will be null and its POS will be the POS attached to it

Prefix will be the first token and its POS will be the POS attached to it

Immediate token will be the first next token to the tagged token and its POS will be the POS attached to it

Next token will be the token next to immediate token and its POS will be the POS attached to it

End if

End if

End for

Figure 6.3 Algorithm for Feature Extractor

6.4.4 Preprocessing the Data Using Weka

After the features are extracted the next step is to preprocess the data using Weka filtering tools. The *StringToWordVector* conversion filtering tool is used. It has the same purpose to that of its use for text categorization. The purpose of the filtering tool is to convert string attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings. The set of words (attributes) is determined by the first batch filtered (typically training data).

6.4.5 Training a Classifier Model

After the data is preprocessed it is passed to the classifier learners in Weka. The explorer or the command line can be used. Weka classifiers can also be used from Java code which we use for the implementation of the prototype system. The decision tree, naïve Bayes and SMO classifier learning algorithms are selected among the different available machine learning algorithms that Weka supports. These algorithms are selected after considering the possible machine learning algorithms that can work on text data and numeric data. We select the three algorithms to see the performance of the classifier and evaluate their efficiency in predicting the category for the candidate token.

6.4.6 Using Trained Classifier for Information Extraction

The purpose of information extraction is to extract information from unseen news text. Once the classifier model is generated using the training data the next step is to use the trained classifier model to work on the unseen news text. The problem is the trained classifier model doesn't directly apply on the plain news texts. It rather works on the extracted features from the news texts. To use the classifier model for extraction first the candidate texts should be identified and the features of the candidate texts must be extracted using the feature extractor. In order to achieve this there is a candidate text selection component which automatically tag the candidate tokens. After the data is preprocessed and the POS of the tokens is identified and a name stemmer is applied then the candidate text selection will be applied.

Candidate Text Selection

The candidate texts for extraction for this research work are names and numbers. The name of a person is a candidate text for Reporter attribute, place name is for place the infrastructure is built on, the number in the news text is considered as a candidate text for number of users and the for the amount of money spent for the infrastructure development, organization names are used as candidate texts for attribute financial source for infrastructure development and names used for infrastructure are the candidate texts. The task of candidate text selection is to tag all these candidate text accordingly to their attributes. The Gazetteers is used for candidate text selection purpose. The gazetteer which consists of names for different places in Ethiopia, different names that can be used for identification of persons, the different infrastructure names, and the different governmental and nongovernmental organization list is used. The candidate text selection is done using the Gazetteer list which comprises of the different names under consideration.

Feature Extraction

Once the candidate texts are selected and tagged, the feature extractors then extracts all the features from the tagged candidate texts and store it in the database for later processing by Weka.

Extraction of data

After the features are extracted and preprocessed using weka the trained classifier model is used to predict the category of the candidate text. Among the candidate texts those with the token category of the predefined attributes will be stored in the database and others which are not under the category of the predefined attributes will be discarded.

6.5 Post Processing of the Data

The post processing handles formatting of the data after the extraction. The attributes are formatted according to the predefined order of the table formation. In this research work the six attributes that are extracted are stored in the table as follows.

6.6 The Prototype User Interface

Based on the designed model the prototype system is developed to evaluate the performance of the proposed model. The following figure shows the ATIE system after extracting the data from the infrastructure news text. The extracted data from the news texts are then stored in the database for further use by users or any other application.

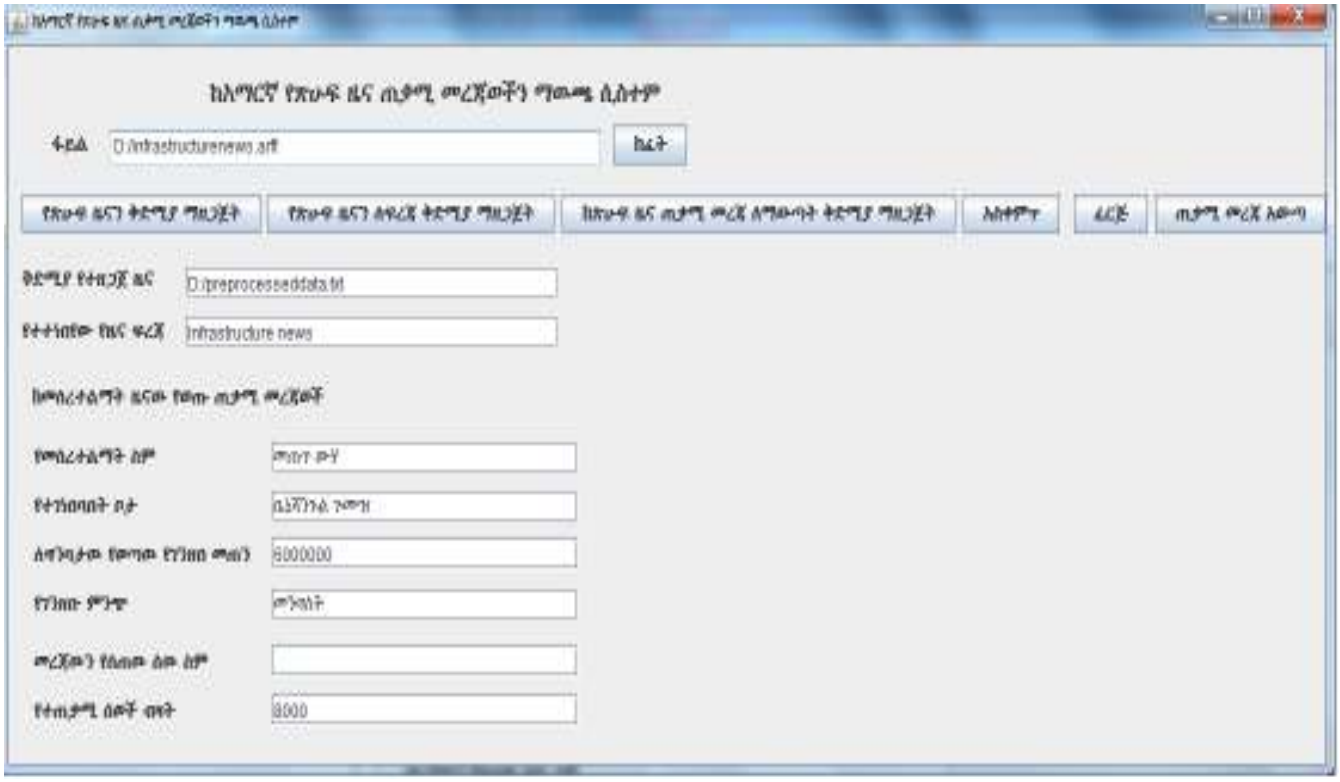

Figure 6.4 Sample Prototype System after the Data Is Extracted From the News Text

For the development of the prototype ATIE system Java programming language, Ling pipe and Weka open source machine learning algorithms are employed. The different subcomponents of the text categorization and information extraction components, which are mostly language specific issues are developed using java based on the designed algorithms. The main reason that the java programming language is used is that machine learning algorithms of Weka are developed using java which can easily be imported and used in the java environment. Among the different machine learning algorithms that Weka supports the one that can be applied for textual and numeric data are selected and used.

To use the prototype the news texts can be opened from the location where it is stored as a file and by using "ከፊት" button. After the file is opened will be first be preprocessed using the button "የጽሁፍ ዜናን ቅድሚያ ማዘጋጀት ", which will tokenize the news text, normalize the different characters, stem the prefix from the number, and normalize the number representation in the news text. After the news text is preprocessed the output can be saved in a selected location by using the "አስቀምጥ" button. To make the preprocessed news text ready for text categorization another button which is the "የጽሁፍ ዜናን ለፍረጃ ቅድሚያ ማዘጋጀት " will be used. When the button is pressed the different subcomponent under the text categorization component will be called and applied on the preprocessed text to select features, which will be used for prediction purpose later. After the news text is preprocessed for categorization the "ፈርጅ" button is used to categorize the news text as one of the three predefined categories. The "ከጽሁፍ ዜና ጠቃሚ መረጃ ለማዉጣት ቅድሚያ ማዘጋጀት "button is used to preprocess the news text for information extraction. The different subcomponents under learning and extraction component are applied on the preprocessed infrastructure news text, which will select the candidate texts and select those candidate texts which will be used as attributes for the predefined database slot. After it is preprocessed, the "ጠቃሚ መረጃ አዉጣ " button can be used to extract the information that can be used as an attribute of the six predefined attributes. The candidate texts which are categorized as one of the six predefined attributes will be stored in the database according to their attribute value as the main purpose of the research is to extract and store the data in the database for further use and management of the data by the users or applications.

Summary

The ATIE prototype system which is developed in this research work has four main components which are interconnected with each other to extract information from infrastructure news text. The document preprocessing component handles language specific issues, text categorization component classify the news text as one of the predefined categories, the learning and extraction component train a classifier and use the classifier for extract information and post processing format the extract data and store them in the database. The different algorithms and subcomponents under the four main components are also presented.

CHAPTER 7

7. EXPERIMENT

In this chapter the test data set, the evaluation metrics and the evaluation result of the proposed ATIE system will be presented. The standard methods that are used to evaluate a classifier model are employed to evaluate the performance of our system.

7.1 Experimental Procedures

IE system is a domain specific activity which needs to be applied on a specific domain which share similar patterns. Using IE system for a different domain for which they are not developed for will not function correctly. Therefore, the following tasks are carried out in order to evaluate the ATIE system we proposed.

7.1.1 Data Collection

Among the different news texts which exist in ENA's database the economy news texts are used for training and testing purpose of our system. As we are evaluating two different components of ATIE i.e. text categorization and IE we prepare two different data sets, which will be used for training and testing of both components. For text categorization 1200 news texts which comprises of 400 news of infrastructure subcategory, 400 news of investment subcategory, and 400 news of subcategory others is used as a dataset for training and testing of the text categorization component of the ATIE system. For the information extraction component 300 news texts from the infrastructure subcategory is used as a dataset for training and testing purpose.

Test Data Set Preparation

Weka support four different evaluation techniques which are using training set, using supplied test set, using fold cross validation and using percentage split. The training set option uses the training set for training the classifier model and evaluating the classifier by using the same data while the supplied test set uses independent test set for performance evaluation. In cross validation option the dataset is randomly reordered and then split into n folds of equal size. In each iterations, one fold is used for testing and the other n-1 folds are used for training the

classifier. The test results of each iterations are then collected and averaged over all folds which give the cross-validation estimate of the accuracy. If the fold is 10 the data set will be divided into 10 equal datasets and at a time one of the dataset is used as test dataset and the other 9 are used as a training dataset it will continue 10 times by making each of the 10 dataset as testing dataset. The output is then added together and the average will be the performance result. The percentage split on the other hand allows splitting the dataset into training and test set randomly by using the specified percentage by the user. [42].

Among the above different evaluation options in Weka we use the 10 fold stratified cross validation technique for the text categorization and IE as it is good to get a better result out of the classifier. Therefore the test data set are not prepared manually as they are automatically created by the Weka itself.

7.2 Performance Evaluation

Since the Weka machine learning algorithms are used for training and testing the classifier model we use the standard evaluation techniques used for evaluating the performance of classifier algorithms. The accuracy of a classifier is measured by using classifier's performance on a test data. It is the percentage of test instances that are correctly classified by the classifier and those that are classified incorrectly. The category assignment of a classifier is evaluated using the confusion matrix which analyzes how well a classifier recognizes instances of different classes and also clearly indicate where the confusion lies while predicting the category. The Weka itself automatically generates the different experimental results.

7.2.1 Evaluation of Text Categorization Component

The text categorization component is responsible for categorizing any economy news text as Infrastructure, Investment or others category. Among the different machine learning classifier algorithms SMO, Decision tree and Naïve Bayes classifier are used for evaluating their performance. The dataset contains 1200 instances which comprises 400 for each subcategory under economy news.

7.2.1.1 Evaluation of Text Categorization Component Using Decision Tree Classifier

Decision tree contain different induction algorithms. The induction algorithms are *ADTree, BFTree, DecisionStamp, Id3, J48, LMT, MSP, NBTree, RandomForest, RandomTree, REPTree*, and *Simplecast*. All these tree induction algorithms are not used for training and testing the classifier for text categorization component. *ADTree* cannot handle non-binary class (the experimental data has nominal classes), Id3 cannot handle numeric attributes. *MSP* cannot handle nominal class. Most of the decision tree classifiers require a large amount of time and memory for training and evaluating the classifier. J48 is the only classifier which requires less memory and time and we test the dataset using J48 and the following is the output result.

Experimental Result

Correctly Classified Instances 1007 83.9167 %

Incorrectly Classified Instances 193 16.0833 %

Confusion matrix

Table 7.1 Confusion Matrix for Text Categorization Component Using Decision Tree

Infrastructure	Investment	Others	Classified as
331	14	55	Infrastructure
6	377	17	Investment
77	24	299	Others

The similarity between the news sources increases the confusion for the classifier in predicting the category of the news document. For example in the first row of the confusion matrix table among the 400 Infrastructure news 331 of them are correctly classified as infrastructure news while the other 77 news confuses the classifier and predict 14 of them as Investments news and 55 of them as others news. The same is true for the second and third row which also show the confusion of the classifier in predicting the Investment and others news category. The confusion is minimum while predicting the Investment news by correctly classifying 377 of the investment

and confused only in predicting 33 of them as infrastructure and others news category. This is due to the reason that the nature of the news text has a very similar attributes that confuse the prediction of the classifier. The confusion matrix is later used to calculate the detail accuracy of the classifier.

Detailed Accuracy by Class

Table 7. 2 Detailed Accuracy by Class for Text Categorization Component Using Decision Tree

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.828	0.104	0.8	0.828	0.813	0.868	Infrastructure
0.943	0.048	0.908	0.943	0.925	0.954	Investment
0.748	0.09	0.806	0.748	0.776	0.836	Others
0.839	0.08	0.838	0.839	0.838	0.886	Weighted Avg.

The detail accuracy by class result of the text categorization component using decision tree describes the TP (True Positive), FP(False Positive), precision, recall, F-measure and Roc Area for each subcategory of the news text the classifier predicts. TP Rate indicates the rate at which the classifier correctly predicts the category of the news text. For example, in the first row the table TP rate of Infrastructure news text is 0.828, which means the rate of the classifier in correctly identifying the news text is 0.828 while is the FP rate is 0.104. As it is described in chapter 2 the precision and recall are also presented. The ROC Area indicates the curves in identifying the true positive over the false positive instances. Among the three categories the ROC Area of Investment news is higher as the number of true positives is higher than the other subcategories and the numbers of false positive instances are lower than other subcategories. The J48 classifier in the decision tree is the better algorithm that can perform on limited memory and requires less processing time while others require a tremendous amount of time and memory which make it difficult to evaluate them. As it is claimed in the work of Yohannes Afework [33] LMT decision tree classifier is better than all the other classifiers in decision tree however it is difficult to evaluate it as it require large amount of memory and time.

7.2.1.2 Evaluation of Text Categorization Component Using Naïve Bayes

The other classifier algorithm that Weka support and which can be applied on textual data is Naïve Bayes. Like decision tree Bayes also comprises different classifier algorithms and among these algorithms the one which perform better is DMNBtext classifier. It is evaluated on the same data set using the same test option which has the following experimental result.

Experimental result

Correctly Classified Instances	1114	92.8333 %
Incorrectly Classified Instances	86	7.1667 %

The accuracy of the DMNBtext classifier is better than the decision tree classifier by correctly predicting the category of 1114 news texts. The confusion matrix also show that the Naïve Bayes algorithm perform better than that of the decision tree which reduces the confusion of the classifier in predicting the category for the news texts.

The confusion matrix

Table 7. 3 Confusion Matrix for Text Categorization Component Using Naïve Bayes

Infrastructure	Investment	Others	Classified as
372	7	21	Infrastructure
5	383	12	Investment
29	12	359	Others

The confusion in categorizing the news text is better than that of the decision tree. For example, in the first row of the confusion matrix table among the 400 Infrastructure news texts, 372 of them are classified correctly while the classifier predicts 7 of them as Investment and 21 of them as others. The same works for the remaining two line of the table. The increase in the correctly classified news texts will minimize the confusion. Since the confusion is less and the accuracy of the classifier in correctly predicting the category of the news text is higher than Decision tree , the precision, recall, TP rate values are increased while using Naïve Bayes algorithm.

Detailed Accuracy by Class

Table 7. 4 Detailed Accuracy by Class for Text Categorization Component Using Naïve Bayes

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.93	0.043	0.916	0.93	0.923	0.976	Infrastructure
0.958	0.024	0.953	0.958	0.955	0.986	Investment
0.898	0.041	0.916	0.898	0.907	0.983	Others
0.928	0.928	0.036	0.928	0.928	0.928	Weighted Avg

7.2.1.3 Evaluation of Text Categorization Component Using SMO

The other classifier algorithm that Weka support and which can be applied on textual data is SMO. The experimental results while using SMO as a classifier are presented as follows

Experimental result

Correctly Classified Instances	1099	91.5833 %
Incorrectly Classified Instances	101	8.4167 %

The Confusion Matrix

Table 7.5 Confusion Matrix for Text Categorization Component Using SMO

Infrastructure	Investment	Others	Classified as
355	8	37	Infrastructure
6	386	8	Investment
34	8	358	Others

Detailed Accuracy by Class

Table 7.6 Detail Accuracy by Class for Text Categorization Component Using SMO

TP Rate	FP Rate	Precision	Recall	F-Measure	Roc Area	Class
0.888	0.05	0.899	0.888	0.893	0.927	Infrastructure
0.965	0.02	0.96	0.965	0.963	0.979	Investment
0.895	0.056	0.888	0.895	0.892	0.931	Others
0.916	0.042	0.916	0.916	0.916	0.946	Weighted Avg.

7.2.1.4 Comparison of the Performance Classifiers Used for Text Categorization

Among the three classifiers that we used to evaluate the text categorization component of our ATIE system the Naïve Bayes algorithm perform better than others. The experimental result of SMO is also very similar to that of Naïve Bayes.. The Naïve Bayes algorithm doesn't only perform better in correctly classifying the news text but also require less time and memory to execute the training and prediction. From the experimental result, we can conclude that when there is an attribute similarity among the classes that are considered for text categorization the Naïve Bayes will perform better.

7.2.2 Evaluation of IE component

The IE component is the crucial component of this research work which uses different features for predicting the categories of the candidate texts. The main focus of the evaluation of IE component is to see the role of the different features that are considered and their effect in efficiently categorizing the candidate tokens as one of the predefined attributes. Four different scenarios are considered for the evaluation to see the role of the different features in the prediction of the category process. The different features that are considered in this research work are POS of the candidate token, POS of the prefix, POS of the previous token, POS of the immediate next token, POS of the next token, candidate token category, prefix token, previous token, immediate next token, and next token.

Scenario 1: Using all the features

Scenario 2: Using all the features except the POS tags

Scenario 3: Using all the features except the token category

Scenario 4: Using all the features except token category and POS tags

For the above all the scenarios the Decision Tree, SMO and Naïve Bayes algorithms are used for the experiment. The experimental result of the algorithm which performs better than others in each of the considered scenarios is presented. The same testing option which is used for text categorization component, which is 10 fold stratified cross validation is used.

Scenario 1:

Among the three algorithms selected for experiment the SMO classifier performs better than the other classifiers by correctly classifying 94.6% of the instances correctly while the naïve bayes perform by correctly classifying 93.6 and J48 decision tree algorithm by performing 90.6%. All the instances that are used for training and testing the IE component are 1422 which includes the different instances for the six predefined attributes for extraction and the other candidate texts which are names and numbers but not considered as a fact for extraction. The detailed experimental result of SMO algorithm is presented as follows.

Experimental result

Correctly Classified Instances	1345	94.5851 %
Incorrectly Classified Instances	77	5.4149 %

Confusion matrix

Table 7. 7 Confusion Matrix for IE Component for Scenario 1 Using SMO

Infrastructure Name	Place	Money Spent	Reporter	Number of Users	Financial Source	Category
268	0	0	0	0	0	Infrastructure name
0	253	0	2	4	0	Place
0	0	249	0	0	1	Money spent
0	12	0	121	0	0	Reporter
0	0	0	0	172	0	Number of users
0	0	0	0	0	126	Financial source

From the confusion matrix table we can see that the confusion in predicting the different categories considered in IE is very minimal.

Detailed accuracy by class

Table 7. 8 Detailed Accuracy by Class for IE Component for Scenario 1 Using SMO

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.996	0	1	0.996	0.998	1	Infrastructure Name
0.996	0.003	0.985	0.996	0.991	0.998	Place
0.977	0.003	0.984	0.977	0.981	0.997	Money Spent
0.956	0.007	0.95	0.956	0.953	0.994	Reporter
0.91	0.012	0.883	0.91	0.896	0.985	Number of users
0.955	0.014	0.875	0.955	0.913	0.991	Financial Source
0.946	0.006	0.942	0.946	0.943	0.992	Weighted Avg.

Scenario 2

Among the three algorithms for scenario 2 the Naïve Bayes algorithm performs a little bit better than SMO. Naïve Bayes correctly classified 94.86 while SMO perform 94.58 % and J48 decision tree algorithm correctly classifies 85.72%.

Correctly Classified Instances 1349 94.8664 %

Incorrectly Classified Instances 73 5.1336 %

Confusion Matrix

Table 7. 9 Confusion Matrix for IE Component for Scenario 2 Using Naïve Bayes

Infrastructure Name	Place	Money Spent	Reporter	Number of Users	Financial Source	Category
250	0	0	0	0	0	Infrastructure name
0	269	0	0	0	0	Place
0	0	257	0	2	0	Money spent
0	0	5	180	0	0	Reporter
0	0	0	0	128	0	Number of users
0	0	0	0	0	132	Financial source

Detailed accuracy by class

Table 7. 10 Detailed Accuracy by Class for IE Component for Scenario 2 Using Naïve Bayes

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	Infrastructure Name
1	0.003	0.985	1	0.993	0.999	Place
0.992	0.015	0.935	0.992	0.963	0.999	Money Spent
1	0.006	0.957	1	0.978	0.997	Reporter
0.962	0.021	0.826	0.962	0.889	0.995	Number of users
1	0.012	0.898	1	0.946	0.995	Financial Source
0.949	0.007	0.95	0.949	0.944	0.994	Weighted Avg.

Scenario 3

Among the three algorithms for scenario 3 the SMO algorithm once again performs better than others. The SMO classified the 90.6% of the dataset correctly while the Naïve Bayes perform 90.6% and J48 decision tree algorithm performs 85.8%.

Correctly Classified Instances 1289 90.647 %

Incorrectly Classified Instances 133 9.353 %

Confusion matrix

Table 7. 11 Confusion Matrix for IE Component for Scenario 3 Using SMO

Infrastructure Name	Place	Money Spent	Reporter	Number of Users	Financial Source	Category
0	0	0	0	0	0	Infrastructure name
8	252	0	0	0	9	Place
1	0	253	0	5	0	Money spent
0	0	0	173	0	0	Reporter
0	0	12	0	121	0	Number of users
0	4	6	0	0	116	Financial source

Detailed accuracy by class

Table 7. 12 Detail Accuracy by Class for IE Component for Scenario 3 Using SMO

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.92	0.022	0.898	0.92	0.909	0.973	Infrastructure Name
0.937	0.013	0.944	0.937	0.94	0.983	Place
0.977	0.003	0.984	0.977	0.981	0.996	Money Spent
0.961	0.006	0.956	0.961	0.958	0.996	Reporter
0.91	0.013	0.877	0.91	0.893	0.978	Number of users
0.879	0.02	0.817	0.879	0.847	0.947	Financial Source
0.906	0.013	0.901	0.906	0.903	0.975	Weighted Avg.

Scenario 4

Among the three algorithms for scenario 4 the SMO algorithm performs better than others. SMO classifies 90.64% of the dataset correctly while Naïve Bayes and J48 decision tree algorithm classifies 89.5218 % and 85.7243 % respectively. The detailed experimental result of the SMO algorithm is presented as follows

Experimental result

Correctly Classified Instances 1289 90.647 %

Incorrectly Classified Instances 133 9.353 %

Confusion matrix

Table 7. 13 Confusion Matrix for IE Component for Scenario 4 Using SMO

Infrastructure Name	Place	Money Spent	Reporter	Number of Users	Financial Source	Category
232	3	0	0	0	0	Infrastructure name
2	252	0	0	0	0	Place
0	0	253	0	6	0	Money spent
0	0	0	173	0	0	Reporter
0	0	0	0	123	0	Number of users
2	1	0	0	0	117	Financial source

Detailed Accuracy by Class

Table 7. 14 Confusion Matrix for IE Component for Scenario 4 Using SMO

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.928	0.01	0.951	0.928	0.939	0.98	Infrastructure Name
0.937	0.008	0.966	0.937	0.951	0.982	Place
0.977	0.003	0.984	0.977	0.981	0.996	Money Spent
0.961	0.008	0.945	0.961	0.953	0.996	Reporter
0.925	0.012	0.891	0.925	0.908	0.984	Number of users
0.886	0.02	0.818	0.886	0.851	0.953	Financial Source
0.906	0.012	0.904	0.906	0.904	0.973	Weighted Avg.

7.2.2.1 Discussion on the Experimental Results of IE

The performances of the classifier that are used for IE component have more or less similar output. Among all the three classifiers used on the four different scenarios, the SMO perform better than all others except in one scenario which Naïve Bayes very slightly exceeds the performance of SMO. As it can be seen easily from the confusion matrix and the higher value of precision, recall, ROC rate, TP rate in each of the scenarios, the performance of the classifiers on the IE component is higher. Among the different features considered the token category feature plays a crucial role for prediction by minimizing the performance of the classifier by 4% when it is excluded as a feature. The POS has almost no effect in Naïve Bayes and SMO which has the same performance when it is used as a feature as well while it has a significant impact on the J48 decision tree algorithm which exceeds its performance by 5%. The confusion matrix for all the experiment output shows that the confusion in predicting the candidate text in to its exact attribute category is very minimum which sequentially increases the detailed experimental result of the classifier on the data set which contains 1422 instances which are the collection of the six attributes which are the predefined attributes for extraction and instances which are not important for extraction but have a numeric and name value.

The experimental result shows that using the features that are considered in this research work and open source machine learning classifier algorithms for IE can have a good performance compared to the other state of the art machine learning algorithms.

Summary

The experiment is conducted on economy news category obtained from ENA. Three subcategories which are Infrastructure, Investment, and other categories are considered for text categorization and six predefined attributes are used for IE. The experimental result shows that classifier algorithms can be used for IE and perform as that of the other algorithms for information extraction. Among the three different classifier algorithms that are used for experimentation the SMO 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 POS. Since our POS performance is 80% only using another POS which perform better in predicting the POS tag might increase the performance of the prediction. The confusion matrix in each of the four scenarios considered for IE is small.

CHAPTER EIGHT

CONCLUSION, CONTRIBUTION AND RECOMMENDATION

8.1 Conclusion

The wide spread use of computers and the tremendous growth of information and communication technology in all parts of the world increased the amount of available textual information from time to time. Today there are numerous electronic documents on different issues which can help the day to day life of individuals from teenage kids to older ones. However, the availability of huge amount of information makes it difficult to manually search and acquire the required information from the ocean of unstructured data. The text data in local languages is also increasing from time to time. This is also true for Amharic as there is a growth in development and use of different online news papers and contents. To alleviate this problem different research works have been conducted to extract relevant information automatically.

IE is one of the research areas which mainly focus on extracting relevant facts from the abundantly available text data and convert it to the database for easy use and management of the data. There are different researches which are conducted so far on the area on different languages and domains. The purpose of this research work is to contribute on the development of IE system for Amharic news text. Though it is the first research work which is conducted on the extraction of data from Amharic news text we can confidently say that it is promising to develop an IE system using machine learning approach.

In this research work we proposed a generic model for Amharic Text information extraction based on which we designed an IE system which we called it as ATIE (Amharic Text Information Extraction) system. It has four main components and is developed using java and the open source machine learning algorithms in Weka. The components of ATIE model are document preprocessing, text categorization, learning and extraction and post processing. The document preprocessing component which consists of tokenization, character normalization, and number normalization handles the language specific issues. The text categorization component predicts the category for the unseen news texts. It is used to identify the news text which the

learning and extraction component use it as an input. The learning and extraction component uses different features that are extracted by using feature extractor on the candidate text for predicting the category of the candidate text. The feature which are used for extraction are previous token, prefix, immediate next token, next token, candidate token, and the respective POS for the above listed tokens. The post processing component formats the extracted data according to the format of the predefined database slots.

For experimental evaluation we used a 10 fold stratified cross validation technique which is one of the testing option that Weka support. By using 1200 instances of dataset for training and evaluating the text categorization component and 1422 instances for the IE component we got the experimental result which is very high.

8.2 Contribution of the Work

- A generic model is designed for Amharic text information extraction system that uses a machine learning approach is proposed.
- The use of text categorization with IE for identification of relevant documents and extraction of data from those relevant documents is proposed
- Algorithms are developed for language specific issues which can handle number normalization, Name Prefix Separator, Number Prefix Separator
- A feature extractor algorithms which extract features from tagged dataset is designed
- An algorithm for automatic token tagger which tags candidate texts as a name or a number is designed
- A prototype system for Amharic Text Information Extraction from infrastructure news text is developed
- The experiment result shows the use of open source classification algorithm for IE from Amharic news text is promising

8.3 Recommendation

We showed that machine learning classifier algorithms can be used for extracting information from Amharic news text. However, using them for large scale dataset with higher numbers of attributes might not perform well since it increases the complexity of the classifier as IE is a very time taking task which requires different NLP processing tools for extraction. The IE system proposed in this research work is not complete it requires different improvements to be used at the large scale. The following are our recommendations for future works

- Co-reference is not considered in this research work as the news data that are used for training and testing discuss only about a single issue but all the Amharic text are not like that therefore the inclusion of Co-reference in the further study will be vital to develop a large scale IE system.
- For recognizing names the gazetteer, which contains a different names which we try to collect from the news and other source is used. Using an automatic named entity recognition in later stages might minimize the burden of selecting the named entities
- Incorporating a POS tagger which performs better than the one used in this research work might increase the performance of the system
- Incorporating Amharic spell checkers to minimize the spelling problems which mostly happen in the news text might also have an impact as we manually modify the spelling errors as they have impact for named entity recognition.

Reference

[1] Seid Yimam, Amharic Question Answering System for Factoid Questions, A Thesis Submitted to the School of Graduate Studies of the Addis Ababa University in partial fulfillment for the Degree of Master of Science in Computer Science, 2009

[2] Fabio Ciravegna,Claudio Giuliano,Nicholas Kushmerick,Alberto Lavelli and Ion Muslea, Adaptive Text Extraction and Mining (ATEM 2006),11th Conference of the European Chapter of the Association for Computational Linguistics Proceedings of the Workshop on April 4, 2006, Trento, Italy

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

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

[8] Angel janevski, Information extraction from university web pages, unpublished Ms thesis, Graduate School of the University of Kentucky.

[9] Marin Hassel, Evaluation of Automatic Text Summarization - A practical implementation, Licentiate thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden, 2004

[10] Nancy Chinchor, MUC-7 Information Extraction Task Definition, http://www-24.nist.gov/related_projects/muc/proceedings/ie_task.html, last visited March 16, 2010

[11] Jun Ying, Analysis and Comparison of Existent Information Extraction Methods, TU Darmstadt, Knowledge Engineering Group, 2006

[12] Un Yong Nahm and Raymond J. Mooney, A Mutually Beneficial Integration of Data Mining and Information Extraction, In the Proceedings of the Seventeenth National Conference on Artificial Intelligence(AAAI-2000), pp.627-632, Austin, TX, 2001

[13] Information Extraction, http://gate.ac.uk/ie/ , last visited December 29, 2009

[14] Dipanjan Das Andre F.T. Martins, A survey on automatic text summarization, unpublished, Language Technologies Institute Carnegie Mellon University, Nov 21, 2007

[15] 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

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

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

[18] Ying Yu, Xiao-Long Wang, Yi Guan, Information Extraction for Chinese Free Text Based On Pattern Match Combine With Heuristic Information, Proceedings of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002

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

[20] Hyun Chul Lee, Jian Chang Mao, Information extraction by embedding HMM to the induced linguistic feature space, International Conference on Intelligent Computing (ICIC), China 2005

[21] Douglas E. Appelt, David J. Israel, Introduction to Information Extraction Technology, a Tutorial Prepared for IJCAI-99

[22] Dwi H. Widyantoro1 , Yudi Wibisono, Information Extraction for E-Job Marketplace, The 4th International Conference TSSA 2007

[23] Alberto Téllez-Valero, Manuel Montes-y-Gómez, Luis Villaseñor-Pineda, Using Machine Learning for Extracting Information from Natural Disaster News Reports, CICLing 2005: 539-547

[24] Aidan Finn, Nicholas Kushmerick, Information extraction by convergent boundary classification, In AAAI-2004 Workshop on Adaptive Text Extraction and Mining. San Jose, USA, 2004

[25] Jie Tang, Mingcai Hong, Duo Zhang, Bangyong Liang, and Juanzi Li, Information Extraction: Methodologies and Applications, In the book of Emerging Technologies of Text

Mining: Techniques and Applications, Hercules A. Prado and Edilson Ferneda (Ed.), Idea Group Inc., Hershey, USA, 2007

[26] Nicholas Kushmerick , Edward Johnston , Stephen McGuinness, Information extraction by text categorization, In The IJCAI-2001 Workshop on Adaptive Text Extraction and Mining

[27] An De Sitter, Walter Daelemans, In Proceedings Of The ECML/PKDD 2003 Workshop On Adaptive Text Extraction And Mining (ATEM 2003) , Cavtat-Dubronik

[28] Yaoyong Li, Kalina Bontcheva, Hamish Cunningham (2005). SVM Based Learning System for Information Extraction. J. Winkler, M. Niranjan and N. Lawerence (Eds.): Deterministic and Statistical Methods in Machine Learning, LNAI 3635, Springer Verlag, 319-339

[29] Benjamin Rosendfeld, Ronen Feldman, Moshe Fresko, TEG—a hybrid approach to information extraction, Knowledge Information Systems (2005) 00: 1–18

[30] João Cordeiro, Pavel Brazdil, Learning Text Extraction Rules, Without Ignoring Stop Words, PRIS 2004: 128-138

[31] Rattasit Sukhahuta and Dan Smith, Information Extraction for Thai Documents, Proceedings of the fifth international workshop on Information retrieval with Asian languages, 103 – 110, Hong Kong, China ,2000

[32] 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, 4-5 November 2002

[33] Yohannes Afework, Automatic Amharic text categorization, A Thesis Submitted to the School of Graduate Studies of the Addis Ababa University in partial fulfillment for the Degree of Master of Science in Computer Science, 2007

[34] Ibrahim Yassin Hamid , Automatic Information Extraction For Amharic Language Text Using Hidden Markov Model , A Thesis Submitted to Graduate School of Telecommunication and Information Technology in partial fulfillment for the Degree of Master of Science in Information Technology,2009

[35]Amharic, http://www.lonweb.org/link-amharic.htm, Accessed on June 20,2010

[36] ባ የ ይ ማ ም፤ 1987 የ አ ማ ር ኛ ሰ ዋ ሰ ው፤ ት .መ.ማ.ማ.ድ .

[37] Hamish Cunningham, Information extraction automatic, Elsevier Science 18th November 2004

[38] Goncalo Simoes, Helena Galhards, Luisa Coher, Information Extraction tasks: a survey

[39] Yannick Versley, A Modular Toolkit for Coreference Resolution, a hand book on coreference resolution

[40] Melese Tamiru,2009, Automatic Amharic Text Summarization Using Latent Semantic Analysis, A Thesis Submitted to the School of Graduate Studies of the Addis Ababa University in partial fulfillment for the Degree of Master of Science in Computer Science.

[41] Tessema Mindaye, 2007, Design and Implementation of Amharic Search Engine, A Thesis Submitted to the School of Graduate Studies of the Addis Ababa University in partial fulfillment for the Degree of Master of Science in Computer Science.

[42] The Weka Documentation, http://www.cs.waikato.ac.nz/ml/weka/, Accessed on October 22, 2010

[43] J. Han and M. Kamber. Data Mining: Concepts and Techniques (2nd ed.). Morgan Kaufmann Publishers, 2006.

[44] Y. Yang. An Evaluation of Statistical Approaches to Text Categorization. The Netherlands: Kluwer Academic Publishers, 1999.

[45] Summary and statistical report of the 2007 population and housing census, www.csa.gov.et/docs/cen2007_firstdraft.pdf, accessed on November 2, 2010

Appendices

Appendix A: List of Stop Words

ሁሉ	ብዛት	ናት	እንደአስረዱት	የታች
ሁሉም	ብዙ	ናቸው	እንደገና	የውስጥ
ኋላ	ቦታ	አሁን	ወቅት	የጋራ
ሁኔታ	በርካታ	አለ	እንዲሁም	ያ
ሆነ	በሰሞኑ	አስታወቀ	እንጂ	ይታወሳል
ሆኑ	በታች	አስታውቀዋል	እዚህ	ይህ
ሆኖም	በኋላ	አስታውሰዋል	እዚያ	ደግሞ
ሁል	በኩል	እስካሁን	እያንዳንዱ	ድረስ
ሁሉንም	በውስጥ	አሳሰበ	እያንዳንዳቸው	ጋራ
ላይ	በጣም	አሳስበዋል	እያንዳንዷ	ግን
ሌላ	ብቻ	አስፈላጊ	ከ	ገልጿል
ሌሎች	በተለይ	አስገነዘቡ	ከኋላ	ገልጸዋል
ልዩ	በተመለከተ	አስገንዝበዋል	ከላይ	ግዜ
መሆኑ	በተመሳሳይ	አብራርተዋል	ከመካከል	ጥቂት
ማለት	የተለያየ	አበራርተው	ከሰሞኑ	ፊት
ማለቱ	የተለያዩ	አስረድተዋል	ከታች	ደግሞ
መካከል	ተባለ	እስከ	ከውስጥ	ዛሬ
የሚገኙ	ተገለጸ	እባክህ	ከጋራ	ጋር
የሚገኝ	ተገልጿል	እባክሽ	ከፊት	ተናግረዋል
ማድረግ	ተጨማሪ	እባክዎ	ወዘተ	የገለጹት
ማን	ተከናውኗል	አንድ	ወይም	ይገልጻል
ማንም	ችግር	አንጻር	ወደ	ሲሉ
ሰሞኑን	ታች	እስኪደርስ	ዋና	ብለዋል
ሲሆን	ትናንት	እንኳ	ወደፊት	ስለሆነ
ሲል	ነበረች	እስከ	ውስጥ	አቶ
ሲሉ	ነበሩ	እዚሁ	ውጪ	ሆኖም
ስለ	ነበረ	እና	ያለ	መግለጹን
ቢቢሲ	ነው	እንደ	ያሉ	አመልክተዋል
ቢሆን	ነይ	እንደገለጹት	ይገባል	ይናገራሉ
ብለዋል	ነገር	እንደተገለጸው	የኋላ	
ብቻ	ነገሮች	እንደተናገሩት	የሰሞኑ	

Appendix B: List of Titles

አቶ	ክቡር	ሻለቃ
ወ/ሮ	ክብርት	ጀኔራል
ወ/ሪት	ሻምበል	ጀነራል
ዶ/ር	ኮሎኔል	ፕሮፌሰር
ሸህ	አስር አለቃ	ወታደር
ቄስ	አምሳ አለቃ	ኢንጅነር

ድያቆን

ባላምበራስ

ብላቴን ጌታ

ፊታውራሪ

ብላታ

አባ

ደጃዝማች

ኰሎኔል

ሜጀር

ጀነራል

በጅሮንድ

መምህር

ግራዝማች

ብላቴን ጌታ

ባላምባራስ

ሊቀ ጠበብት

ዶክተር

ሻንበል

ነጋድራስ

ኰሎኔል

ልኡል

ራስ

አቡነ

መምህር

አለቃ

ብላታ

ሀኪም

ነጋድራስ

ሀጂ

አርቲስት

አፈ-ጉባኤ

የተከበሩ

አምባሳደር

ኮማንደር

ብርጋድየር ጀኔራል

ሌተናል ኮሌኔል

ሹም

ፕ/ር

አፄ

መቶ አለቃ

ሚስተር

ጠ/ሚ

ሚኒስትር ድኤታ

ብፁእ

ዶክተር

ተመራማሪ

ከንቲባ

ሊቀመንበር

ምክትል

ሳጅን

አ/አለቃ

ከንቲባ

ክቡር

ሎሬት

ሀምሳ አለቃ

አሰልጣኝ

አምበል

ኡስታዝ

ኢንስትራክተር

ወ/ሮ

ኢንጂነር

ሰአሊ

ፒያኒስት

ሚ/ር

ጠ/ሚኒስትር

ጠ/ ሚኒስትር

ፕሬዝዳንት

ፕረዝዳንት

ፕሬዚዳንት

ፕሬዚደንት

ፕረዝደንት

ካፒቴን

ፓትሪያርክ

ፕ/ት

አፈ ጉባኤ

ማእድንና ኢነርጂ ሚኒስትር

ወይዘሮ

ጠቅላይ ሚኒስትር

ሸክ

ዋና ዳይሬክተር

ዳይሬክተር

ኢንስፔክተር

Appendix C: List of Infrastructure Names

መንገድ	ጤና ጣቢያ	ማሰልጠኛ ተቋም
መብራት	ሆስፒታል	አስፋልት መንገድ
ስልክ	ትምህርት ቤት	ጠጠር መንገድ
መጠጥ ዉሃ	አንደኛ ደረጃ ትምህርት ቤት	ኮብል ስቶን መንገድ
ንጹህ መጠጥ ዉሃ	ሁለተኛ ደረጃ ትምህርት ቤ ት	ሞባይል ኔትወርክ
ጤና ተቋማት	ዩኒቨርሲቲ	አዉቶማቲክ ስልክ
ክሊኒክ	ኮሌጅ	ዋይርለስ ስልክ

Declaration

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

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____