JIMMA UNIVERSITY

JIMMMA INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING

EXTRACTING RELATIONS BETWEEN AMHARIC NAMED ENTITIES USING A HYBRID APPROACH

By: Selomon Getnet

THESIS SUBMITTED TO FACULTY OF COMPUTING JIMMA INSTITUTE OF TECHNOLOGY IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTERS OF SCIENCE IN INFORMATION TECHNOLOGY

JIMMA UNIVERSITY

JIMMA INSTITUTE OF TECHNOLOGY

FACULTY OF COMPUTING

EXTRACTING RELATIONS BETWEEN AMHARIC NAMED ENTITIES USING A HYBRID APPROACH

By: Selomon Getnet

ADVISOR:

Getachew Mamo (PhD)

CO-ADVISOR:

Mr. Tesu Mekonen (Msc)

APPROVED BY

Dr. Getachew Mamo Advisor ______________

Mr. Tesfu Mekonen Co-Advisor _____________

January 22, 2020

Declaration

I declared that this thesis is my original work and has not been submitted as a partial requirement for a degree in any university.

Selomon Getnet

January 22, 2020

Acknowledgment

First and foremost i would like to thank the Almighty God and st. marry for giving me the strength to do this thesis and making every thing possible. I would extend my gratitude to my advisor Dr. Getachew Mamo for his guidance, feedback and constructive comments during this work. I would also like to thank my co-advisor Mr. Tesfu Mekonen for his support, encouragement and continues advice through out this work.

I would also thank my friends and my colleagues especially Dawud yimer and Gashaw Demlew for their support, suggestions and sharing ideas.
Finally i would like to thank my family for their endless support.

Contents

List of Figures

List of Tables

Abbreviation

ACE: Automatic Content Extraction

AFF: Affilation

ANERE: Amharic Named Entity Relation Extraction

COM: Company

CoNLL: Conference on Natural Language Learning

CRF: Conditional Random Field

CSV: Comma Separated Value

DARP: Defense Advanced Research Projects Agency

EDT: Entity Detection and Tracking

EL: Entity Linking

GPE: GeoPoletical Entity

IE: Information Extraction

IR: Information Retrieval

KB: Knowledge Base

LOC: Location

LP: Label Propagation

ML: Machine Learning

MUC:Message Understanding Conference

NE: Named Entity

NER: Named entity Relation

NLP: Natural Language Processing

ORG: Organization

PERS: PERSON

POS: Part Of Speach

RDC: Relation Detection and Characterization

SVM: Support Vector Machine

SOC: Social

Abstract

Currently the number of electronic data is increasing than ever before and we can find high frequency of named entities in electronic texts. Named entity relation extraction is the process of finding the relation between two named entities from input text, which is a foundation of semantic networks, ontology design and widely used in information retrieval and machine translation as well as question and answering systems. In this study we develop a hybrid approach by combining a machine learning approach using Support vector machine (SVM) and set of rules. We first used the classifier to predict relations found between named entities. And then to improve the result which is obtained from the machine learning component we used set of rules. Precision, recall and f-measure are used to measure the performance of our proposed system. We have used a total of 764 annotated sentences for training and testing purpose. Our testing is conducted for specific relationship types separately and the highest precision value achieved in this work is 94% for ዋና- ከተማ, the highest recall is also 96% for ጠቅላይ- ሚኒስቴር and the highest f-score is 92% for ዋና- ከተማ. To measure the overall performance of the system we take the average value and it gives us 80%, 81% and 83% of precision, recall and f-score value respectively.

Key words: Relation extraction, Amharic named entity relation extraction, Named entities, Support vector machine, Hybrid approach for relation extraction

Chapter One

1 Introduction

Named entity relation extraction is one of the main tasks of information extraction. It takes semi structured or unstructured text as an input and its task is to identify various semantic relations between entities from text. For example the sentence የአፍጋኒስታን ፕሬዚዳንት አሽራፍ ጋኒ በዛሬው እለት አዲስ አበባ ገብተዋል ። ("Afghanistan president Ashraf Ghani arrives in Addis Ababa today.") carries the semantic relationship ፕሬዚዳንት ("president") between named entities አሽራፍ ጋኒ (Ashraf Ghani)(PERSON) and አፍጋኒስታን (Afghanistan) (GPE). Extracting relation that describes any semantic interaction found between named entities is a very important research topic in the area of information extraction. MUC-7/MET-2 [2] gives a specific definition of named entities on the level of entity extraction as Named Entities (NE) is proper names and quantities of interest, Person, organization, and location names were marked as well as dates, times, percentages, and monetary amounts. The recognition of these entities is basic and first task for building semantic analysis and information extraction system. Named entity extraction is an information extraction task aimed at identifying and classifying words of a sentence, a paragraph or a document into predefined categories of named entities. The idea of Named Entity Recognition (NER) is identifying named entities like people, place, date, number and etc. The first step in named entity recognition is the identification of proper nouns from a text and the second task is the classification of these proper nouns in to any one of the classes like person name, organization name, place name etc.

We can define Relation Extraction as the process of recognizing the type of relation that connects two or more Named Entities. As stated in [4] first, the concept of relation extraction was introduced as part of the Template Element Task, one of the information extraction tasks in the Sixth Message Understanding Conference (MUC-6)[68]. MUC-7 added a Template Relation Task, with three relations. Following MUC, the Automatic Content Extraction (ACE) meetings [64] are pursuing information extraction. In the ACE Program, Relation Detection and Characterization (RDC) were introduced as a task in 2002.

Named entity relation extraction is a significant research topic in the field of information extraction, and aims at finding various semantic relations between named entities [6]. This constitutes very important move toward natural language processing (NLP) applications. This type of information is enables the task of discovering a useful relationship or interaction between entities [7]. A relation among Named Entities can be either introduced directly through words from a context or expressed implicitly from a context of a sentence.

The extraction of relations between named entities received a high attention because Named Entity Relations are a foundation of semantic networks, ontology and the semantic Web, and are widely used in information retrieval and machine translation, as well as automatic question and answering systems [6]. In fact, the Named Entities relations extraction can be exploited to extract more precise and correct answers. For instance if we take the example "where was Alemu born?" the expected answer will be Alemu was born in Jimma. The relational triple is born-in (Person, Location), where Person and Location are the Named Entities. So to give the answer like the above quires we have to analyze relevant documents to collect the necessary information. Indeed, there is a growing need to automatically extract semantic knowledge from texts. Thus, we have to go beyond the detection of named entities and try to extract relation between them.

Therefore, several studies on NE recognition have already been performed in many languages, such as English[58, 45, 60], Arabic[63, 65, 66], and Chinese[61, 62]. Relation extraction from Amharic named entities has not received a significant concern when it compare with English, Arabic and chines languages. Some named entities recognition systems have been done for Amharic language. From those proposed systems [8] based on hand crafted rules, called rule based approach using gazetteers and [9] based on supervised machine learning approach.

Several methods have been proposed to extract semantic relation between named entities. These methods can be classified as rule based, machine learning and hybrid approach. Rule based approach contains set of hand written rules. Rules are written by the language experts. So for this approach human experts are required. The rule-based method offers a significant analysis of the context for each Named Entities and its relations with the other Named Entities. To extract the relation between named entities a noticeable effort is required to write down all the rules for discovering relations between Named Entities.

To fully automate the relation extraction between named entities machine learning approach has been used. This approach includes supervised, semi supervised and un-supervised techniques. Supervised technique requires a fully labeled corpus. The most often used supervised techniques include support vector machine (SVM), conditional random field (CRF), decision tree and maximum entropy model (ME)[7].

The hybrid approach uses both rule based and machine learning methods. So in the hybrid approach we combine any of the two methods in order to improve the performance of the extraction of relation between named entities. Different studies were developed using hybrid approach for English [25, 67], Arabic [7, 57] and other languages. [25] Used a hybrid approach which combines supervised and rule based approaches to extract relationships from stories. They reported 87% and 79.7% of precision and recall respectively. [25] provide an improvement in precision and recall over [48] which uses a supervised machine learning algorithm using Support vector machine(SVM) and achieved precision of 70.62 and recall of 78.32. The work presented in [7] which is done for Arabic using a hybrid approach achives f score of 75.22% and it outperformed both the rule based approach [10] by 12% and the machine learning approach[68] by 9% in terms of F-score. However there is no hybrid approach developed for Amharic language to extract relations between named entities. So based on the benefit that we get from this method we propose our system which aimed to extract relations between Amharic named entities using a hybrid approach. In this case we are using rules mainly to improve the quality and accuracy of our system output by writing some rules.

1.1 Motivation

Amharic is the second Most spoken semitic language in the world, next to Arabic and the official working language of the federal democratic republic of Ethiopia[71]. A number of literature works, newspapers, magazines and education resources are published and available in Amharic language. Most of the official documents in governmental and private sectors in Ethiopia are written in Amharic[70]. Hence, above all these facts initiate us to do this research.

Many named entity relation extraction systems have been done in many foreign languages such as English[25, 26, 67], Arabic [7, 10, 57], Chines [6, 14]. When compared with other foreign languages Amharic is under resourced language and difficult to find resources related to information extraction. Even though there is a growth of electronic Amharic documents having named entities, there is no any system that extracts relations found between Amharic named entities which contributes for many natural language processing and information extraction tasks. There fore this research contributes a lot in Amharic named entity relation extraction and used as an input for different NLP and IE applications.

Generally, considering the above mentioned issues and the advantage and the application areas of extracting relations between named entities in many NLP applications and information extraction tasks motivated us to do this research.

1.2 Statement of the Problem

Currently the number of Amharic electronic data is increasing than ever before and we can find the high frequency of Amharic named entities in electronic texts. Amharic is the language with rich and complex morphological structures and being the working language of federal government of Ethiopia. A lot of valuable information is being published in Amharic currently and we can find high frequency of Amharic named entities in electronic Amharic documents. Amharic is written with a version of the Ge'ez script known as ፊደል (Fidel) and has its own unique grammar, syntax, character (Fidel) representation and statement formation and spoken by a large number of population. Amharic has a unique features when compared with other languages like English. As it stated in [70] unlike the English which is SVO (Subject, Verb, Object) the Amharic clause order is SOV (Subject, Object, Verb). To show this in example, The Amharic sentence " አበበ በሶ በላ ", አበበ is subject, በሶ is an object and በላ is verb and while the English "Abebe ate besso" , Abebe is subject, ate is a verb and besso is object[70].

According to [11] Amharic is the second most spoken sematic language next to Arabic and the second largest language in Ethiopia(after oromifa, a Cushitic language) and possibly one of the five largest languages on the African continent. Despite having large number of speakers, Amharic is one of under resourced language. And there is no any system for extracting relations between

Amharic named entities that could be contributes for advance researches in natural language processing (NLP) and information extraction systems like ontology design, question answering, machine translation and any other systems.

Named entity recognition can be considered the first step towards semantic analysis of texts and a crucial subtask of information extraction systems. But named entities recognition is only the first step for full language processing. If we want to go beyond the detection of entities, a natural step is establishing semantic relations between these entities. But the relations between these entities are not enough represented in the used resources.

There is a growing need to automatically extract semantic relation from Amharic named entities. As it is stated in [6, 10] Named entity relations are a foundation of semantic networks, ontology and the semantic Web, and are widely used in information retrieval and machine translation, as well as automatic question and answering systems and text summerization. Different researchers propose different methods for extracting relations between named entities. The system developed for extracting relations between named entities of one language that works effectively cannot be work for other language with the same accuracy and efficiency or may not work at all. This is because the relation extraction system between named entities has to be trained with the nature of the given language. In this case developing an efficient system for extracting relations between Amharic named entities is an important task.

To the best of our knowledge there is no study that used a hybrid approach to extract the relationship between Amharic named entities. So this work is the first work to design a hybrid system to combine the advantages of Machine Learning and rule based approaches.

Therefore, we will address the following research questions:

- How Can we extract the relationship between Amharic named entities?
- What are the best feature sets for improving the performance of Amharic named entity relation extraction?
- What type of data is suitable for supervised relation extraction?

1.3 Objectives

The objectives of the research are stated as general and specific objective as follow.

1.3.1 General objective

The general objective of this research is to develop automatic relation extraction system for Amharic named entities using a hybrid approach.

1.3.2 Specific objective

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

- Conduct literature review on relation extraction to better understanding the state of art of relation extraction between named entities
- Amharic text corpus collection and preparation
- Design a model for automatic Amharic named entity relation extraction
- Identify features and methods that bring better performance for the extraction of relations between Amharic named entities
- Evaluate the performance of the system by using the test data

1.4 Research Methodology

Extraction of relation between named entities is a compound task that is to be done with different components in different steps. In this part we outlines the research method, the method of data collection and tools used in this research. Collecting the corpus from different news source was our initial task. Selecting the appropriate tools and techniques were also the tasks need to be done in all stages.

1.4.1 Literature review

To understand better the extraction of relations between named entities and the state of art approaches of relation extraction different related literature from books, journal articles and internet were reviewed in this work.

1.4.2 Corpus collection and Data preparation

Relevant Amharic corpus is collected and prepared from different sources in order to use it for train the system. It is difficult to find publicly available annotated corpora that contain necessary information for Amharic language. This makes machine learning method especially supervised approach difficult to use since it needs a large number of annotated training data. Since Amharic is under resourced language, we cannot find annotated corpus that is helpful for our task. We are forced to prepare our own data set. We constructed our corpus from different Amharic news resources and annotated it with 10 relation types between named entities.These relations are: የትውልድ - ቦታ (Birth-Place), ጠቅላይ - ሚኒስቴር (Prime-minister), ሥራ- አስኪያጅ (Manager), ፕሬዝዳንት (President), የሚገኘው (Locted-in), መኖሪያ - ቦታ (Live-in), መስራች (Founder), ሃላፊ (Curator), ዋና- ከተማ (Capital-city) and መሪ (Leader). Based on the number of annotated data that we collected for each relation types we realized that some relations like ሥራ- አስኪያጅ (Manager), ሃላፊ (Curator), የትውልድ - ቦታ (Birth-Place), መኖሪያ - ቦታ (Live-in) and መሪ (Leader) has small number of annotation examples and does not allow efficient learning driven extraction. Finally we choose only five relation types: ጠቅላይ - ሚኒስቴር (Prime-minister), ዋና- ከተማ (Capital-city), ፕሬዝዳንት (President), የሚገኘው (Locted-in) and መስራች (Founder). During the preparation of the corpus we will select a sentence that contains at least two named entities because our aim is to extract a binary relationship between named entities.

1.4.3 NLP pre processing

Different NLP preprocessing activates are performed in this step including tokenization, POS tagging, Named entity recognition and dependency tree parsing to produce the training data set. Given a text input the pre-processing module divide in to a sequence of words or tokens and then tag each tokens with pos, their dependency value, and named entity type.

1.4.4 Development tool

Python programming language is used to implement different language specifications algorithms and pre-processing tasks.The reason python was selected is because of it is easy for text processing.

1.5 Scope and Limitation

The focus of the study is extracting relationship between Amharic named entities using a hybrid approach. In order to develop a good model for relation extraction we need to have available NLP components but those NLP components are not available publicly. By assuming such constraints, The scope of the study is limited on determining explicit relations found between Amharic named entities which is found with in the same sentence. Relations expressed implicitly and found between different sentences are not include in this research. Named entities can be name of people,organization, locations, Geo-Poletical Entities(GPE) as well as temporal and numeric expressions. But due to the high frequency of specific named entities in Amharic texts we are focus on extracting relations lies between any pair of the following four Named entities, Location, Organization, Person and GPE .

1.6 Application of Results

As stated in the statement of the problem, extracting relationship between named entities is useful for many areas of Natural language processing (NLP) and information extraction of Amharic language. So that the beneficiaries of this research includes researchers involved (want to be involved) in different NLP and information extraction researches in which it needs relations between Amharic named entities. In addition this automatic extraction of relation between Amharic named entities benefits different users by enabling them to get relevant information quickly for their complex quires because entity relation extraction is very useful for question answering and solving complex quires.

Generally Amharic named entities relation extraction is used for different applications, such as:

- Text summarization
- Question answering
- Ontology learning and semantic networks
- Machine translation

1.7 Thesis Organization

This study contains Six chapters. Chapter one is discussed about the general background of the research work, statement of the problem, the general and specific objectives, scope and limitation, research methodology and finally about application areas of the research. Chapter two is about literature review. In this chapter a detail description about information extraction, machine learning approaches for relation extraction and about Amharic language are presented. Chapter three discusses related works done on relation extraction between named entities using different approaches on different languages. The design and implementation part of this research was presented in chapter four. The over all system architecture and its components are discussed in detail in this chapter. Chapter five describes the experimentation and evaluation of this work. Finally conclusions and recommendations based on the result of the experimentation and future works are presented in chapter six.

Chapter two

2 Literature Review

In the sections below different approaches to relation extraction, sub tasks of information extraction which includes named entity recognition, evaluation metrics and any concepts related with named entity relation extraction are reviewed in order to understand the problem domain and the extent of the work to be done.

2.1 Information Extraction

Now a day there is a rapid growth of textual information available in digital form in the internet and other electronic Medias. A significant part of such information like government documents, legal acts, online news, and social media communication is transmitted in unstructured form and thus it is difficult to search in. This resulted in a growing need for effective and efficient techniques for analyzing free text data and discovering valuable and relevant knowledge from it in the form of structured information. This leads to the concept of information extraction technologies. Information Extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources.

The general goal of information extraction is to discover structured information from unstructured or semi-structured text[20]. The IE tasks may vary in detail and reliability, but two subtasks are very common and closely related: named entity recognition and relation extraction. Named entity recognition identifies named objects of interest such as person, organizations or locations. Relation extraction involves the identification of appropriate relations among these entities. Examples of the specific relations are employee-of and parent-of. Employee-of relation holds between a particular person and a certain organization and parent-of holds between a father and his child [12].

Information Extraction has not received as much attention as Information Retrieval (IR) and is often confused with Information Retrieval [13]. Information extraction and information retrieval are two different concepts. IE is differ from IR in which The IR process usually returns a ranked list of documents, where the rank corresponds to the relevance score that the system assigned to the document in response to the query. Whereas The goal of IE is not to rank or select documents, but to extract from the documents relevant facts about pre-specified types of events, entities, or relationships, in order to build more meaningful, rich representations of their semantic content.

The Message Understanding Conference (MUC) [17, 18] and Automatic content Extraction (ACE) [19] program influence the scope of information extraction. Before these two competetions (MUC and ACE) the extraction task were mainly focus on the identification of named entities like person and location names and relations between them from natural language text[16]. Extraction of structured information from a text started gaining much attention when DARPA initiated and funded the Message Understanding Conference in the 90's [20]. Early MUCs defined information extraction as filling a predefined template that contains a set of predefined slots. The message understanding conference MUCs provide a forum for assessing and discussing progress in the field of natural language processing. Each conference is preceded by a formal evaluation of text analysis system that has been developed to perform a shared task, as designed by the government in consultation with evaluation participants from the research community [21].

Automatic Content Extraction (ACE) [19] is an evaluation conducted by NIST to measure the tasks of Entity Detection and Tracking (EDT) and Relation Detection and Characterization (RDC). The Entity Detection task requires that selected types of entities mentioned in the source data be detected, their sense disambiguated, and that selected attributes of these entities be extracted and merged into a unified representation for each entity [22]. As stated in [23] the goal of RDC is to detect and characterize relations of the targeted types between EDT entities. ACE defines the following NE types: PERSON, ORG, LOCATION, FACILITY, GEO POLITICAL ENTITY (GPE), WEAPON. The objective of the ACE program is to develop technology to automatically infer from human language data the entities being mentioned, the relations among these entities that are directly expressed, and the events in which these entities participate.

Structured databases, labeled unstructured data, linguistic tags, etc. are the type of input resources available for extraction. Structured data is the data that can be easily organized. It is simple, clean, analytical and usually stored in databases. Fully structured data follows a predefined schema. A typical example for fully structured data is a relational database system. Unstructured data refers to information that either does not have a predefined data model or identifiable structure. Collection of data from social media is an example for unstructured data.

Usually IE, as many other NLP tasks, can be regarded as a pipeline process, where some kind of information is extracted at each stage. Different types of information extracted are [25]:

- Named Entities (NE)
- Temporal Expressions
- Numerical Values
- Relation between entities

2.2 Sub tasks of information extraction

2.2.1 Named entity recognition

A named entity (NE) is often a word or phrase that represents a specific real-world object. Named entities play a central role in conveying important domain specific information in text, and good named entity recognizers are often required in building practical information extraction systems [8]. The task of named entity recognition is to identify named entities from free-form text and to classify them into a set of predefined types such as person, organization and location. Named Entity Recognition (NER) is one of the major tasks in Natural Language Processing (NLP). It is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money, and percent expressions within a text.

Research on named entity recognition has been promoted by the Message Understanding Conferences (MUCs, 1987-1998), the Conference on Natural Language Learning (CoNLL, 2002-2003), and the Automatic Content Extraction program (ACE, 2002-2005) [27]. At first, Named Entity Recognition (NER) was present as a subtask of MUC-6(Message Understanding Conference) [28]. Throughout the MUC series, the term named entity came to include seven categories; persons, organizations, locations (usually referred to as ENAMEX), temporal expressions, dates (TIMEX), percentages, and monetary expressions (NUMEX). Information extrction makes information easier to locate. This is done by first locating named entities and then categorizing them under different labels. Named entity recognition is probably the most fundamental task in information extraction. Extraction of more complex structures such as relations and events depends on accurate named entity recognition as a preprocessing step. As stated in [29] apart from being a building block in information extraction Named entity recognition has many applications like question answering, information retrieval, machine translation, parsing, meta data for semantic and fast information gathering.

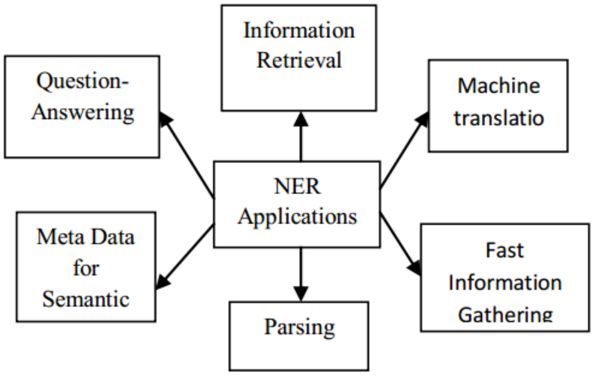

Figure 1: applications of named entity recognition

2.2.2 Relation Extraction

A relation is an aspect or quality that connects two or more things or parts as being, belonging, working together, or as being of the same kind. So, in formal, we can define Relation Extraction as the process of recognizing the type of relation that connects two or more Named Entities. Examples of such entities include: names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. The input is multi-structured data, including structured data (info box form), semi-structured data (tables and lists) and non-structured data (free text). And the output is a set of fact triples extracted from input data. relation extraction (RE) is one of the steps of information extraction. It typically follows named entity recognition and coreference resolution and aims to gather relations between NEs. [32] define relation extraction as:"the task of discovering semantic connections between entities. In text, this usually amounts to examining pairs of entities in a document and determining whether a relation exists between them." Recently it has received more and more attention in many areas like information extraction, ontology construction, and bioinformatics etc.

The concept of relation extraction was first introduced as part of the Template Element Task, one of the information extraction tasks in the Sixth Message Understanding Conference (MUC-6) (Defense Advanced Research Projects Agency, 1995). MUC-7 added a Template Relation Task, with three relations. Following MUC, the Automatic Content Extraction (ACE) meetings (National Institute of Standards and Technology, 2000) are pursuing information extraction.

The relation extraction task identifies various semantic relations such as location, affiliation, revival and so on between entities from text. For example, the sentence የአፍጋኒስታን ፕሬዚዳንት አሽራፍ ጋኒ በዛሬው እለት አዲስ አበባ ገብተዋል ። ("Afghanistan president Ashraf Ghani arrives in Addis Ababa today.") carries the semantic relationship ፕሬዚዳንት ("president") between named entities አሽራፍ ጋኒ (Ashraf Ghani)(PERSON) and አፍጋኒስታን (Afghanistan) (GPE). Many applications in information extraction, natural language understanding, and information retrieval require an understanding of the semantic relations between entities. Extracting semantic relations between entities in natural language text is a crucial step towards natural language understanding applications. A relation is defined in the form of a tuple t = (e1,e2,...,en) where the ei are entities in a predefined relation r within document D. Relations can be found between two entities (binary relation) or more than two entities but Most relation extraction systems focus on extracting binary relations[35].

There are different relation types and [36] present relation types from ACE 2003 and these relation types can be:

- ROLE: relates a person to an organization or a geopolitical entity

  Subtypes: member, owner, affiliate, client, citizen

- PART: generalized containment

  subtypes: subsidiary, physical part-of, set membership

- AT: permanent and transient locations

  subtypes: located, based-in, residence

- SOCIAL: social relations among persons

  subtypes: parent, sibling, spouse, grandparent, associate

Before MUC-7, relations between entities were part of the scenario-specific template outputs of IE evaluations. In order to capture more widely useful relations, MUC-7 introduced the template relation task. 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 [30].

Before starting to extract relations, it is a good idea to determine which words refer to the same "object" in the real world. These objects are called entities. For example, "Barack", "Obama" or "the president" may refer to the entity "Barack Obama". Let's say we extract relations about one of the words above. It would be helpful to combine them as being information about the same person. Figuring out which words, or mentions, refer to the same entity is a process called entity linking [30]. Entity Linking (EL) is a central task in information extraction given a textual passage, identify entity mentions (substrings corresponding to world entities) and link them to the corresponding entry in a given Knowledge Base (KB, e.g. Wikipedia or Freebase)[37].

2.3 Methods to extract semantic relations between named entities

There are three main methods to the design of named entity relation extraction[7].

- Rule based approach
- Machine learning approach
- Hybrid approach

2.3.1 Rule based Approach

Utilize predefined linguistic (syntactic and semantic) rules written manually to extract relationships based on part of speech information. It is very interesting for a restricted domain and has a good quality of analysis. The major drawback of this approach is the disability to perform well in dealing with a wide range or new domain data. This is due to two reasons: rules should be rewritten for different tasks or when the application is enlarged to different domains and finding rules manually is very hard and time-consuming [38].

2.3.2 Machine learning Approach

Machine learning (ML) techniques are widely used as a component of relation extraction methods. Machine learning methods are based on statistical analysis of data to infer general rules. The task of a ML method is either to learn rules from the structure of the underlying data, or to distinguish instances of data from each other. Therefore, the outcome of a ML method is either the learning rules or a model which is used to predict unknown data based on previous seen data [32].
To fully automate the relation extraction task, some research studies have been oriented toward ML methods, including un-supervised, semi-supervised and supervised learning techniques.

2.3.2.1 Supervised machine learning

Supervised learning algorithm requires text corpus in which the entities and their relation types are already known. Such algorithms typically learn to classify new entity pairs into any of the relation types it has already seen, based on some recurring patterns. Supervised learning-based methods have been shown to be effective and perform much better than the other two alternatives. However, their performance much depends on the availability of a large amount of manually labeled high-quality data and annotating large corpora with relation instances is expensive and tedious[39]. Supervised methods based on training set where domain specific examples have been tagged. Such systems automatically learn extractors for relations by using machine learning techniques. The main problem of using these methods is that the development of a suitably tagged corpus can take a lot of time and effort. On the other hand, these systems can be easily adapted to a different domain, provided there is training data [42]. This approach considers relation extraction as a classification task. Support Vector Machines (SVM), Conditional Random Fields (CRF), decision tree and maximum Entropy (MaxEnt) are the most used supervised machine learning techniques.

Supervised techniques for relation extraction can be classified in to two based on the nature of input to the classifier as kernel based methods and feature based methods[35]. Kernel based methods design kernel functions to compute similarities between representation of two relation instances by using for example, shallow parse trees, dependency trees, dependency graph paths etc and employ Support Vector Machine for classification. Feature based methods accepts a set of positive and negative relation examples, syntactic and semantic features are extracted from the text. These extracted features are used to decide weather the entities in the sentence are related or not. Syntactic features extracted from the sentence include the entities themselves, the type of the two entities, word sequence between the entities, number of words between the two entities and path in the parse tree containing the two entities. Semantic features include the path between the two entities in the dependency path. Both the syntactic and the semantic features extracted are given to the classifier in the form of feature vectors for training and classification. Support vector machine or maximum entropy model are used as a classifier.

2.3.2.1.1 Support Vector Machine

Support Vector Machines (SVM) are a supervised machine learning technique. Most relation extraction systems used Support vector machine as a classifier [53]. As [56] cited in their paper an SVM is a vector space based machine learning method, where a goal is to find a decision boundary between two classes that is maximally far from any point in the training data. Basically, a SVMs are binary classifiers. This method works as follows: given a training set, for example a list of annotated sentences. All sentences are transformed into representations such that ML methods can capture properties or features that can best express the interaction pairs. The simplest representation for is a list of words that occurs in the sentences. More complex representations are parse trees obtained from the output of NLP tools which can express the structure and dependencies of words in the sentence. The set of this structured representation and relations, then used as an input for the machine learning classifier, that is, SVM classifier to build a model.
To predict new relations found between named entities from unseen text, every new sentence must be transformed into the same representation as the training sentences, the SVM classifier then used the learned model to extract relations found between pair of named entities.

The reason why we choice Support Vector Machine for this purpose is that there are good implementation of the algorithm available and SVMs are good for binary classification and it achieved better performance in many learning tasks.[69]

2.3.2.2 Unsupervised machine learning

Unsupervised methods use a set of generic patterns to automatically instantiate relation-specific extraction rules and then learn domain-specific extraction rules. The whole process is repeated iteratively. It is also known as self-supervised learning method [42]. Unsupervised learning-based methods normally perform very poorly, though they do not depend on the availability of any manually labeled data .The un-supervised methods make use of massive quantities of unlabeled text. [7].

Unsupervised method do not need fully annotated data or any initial manually selected seeds and are based almost entirely in clustering techniques and similarities between features or context words. According to [9] as they sited in [22] The goal of unsupervised learning is to group data in to clusters. in fact the basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms find out similarity between data in order to determine weather they can be characterized as a cluster.

2.3.2.3 Semi-supervised machine learning

To solve the problems with the unsupervised approach, Some supervised systems also use bootstrapping to make construction of the training data easier. These methods are also sometimes referred to as "weakly supervised information extraction". It uses an initial small set of seeds or a set of hand-constructed extraction patterns to begin the training process. After the occurrences of needed information are found, they are further used for recognition of new patterns. A sample of linguistic patterns or some target relation instances can be used to acquire more basic relations until discovering all the target relations To achieve better balance between human efforts and extraction performance, semi-supervised learning has been drawing more and more attention recently in semantic relation extraction and other NLP applications as well [39].

2.3.3 Hybrid Approach

The two categories of approaches described above can be combined to obtain a mixed approach. Recently, research studies have been oriented toward the use of hybrid approaches because such an approach achieves an enhanced performance that is better than either the rule-based approach or the MLbased approach alone [7]. This approach uses manually handcrafted rules and those extracted from data through Machine Learning (ML)-algorithms [41].Among the systems based on this approach, we can mention the system developed by [7] to extract relations between Arabic Named Entities.The developed system used linguistic modules employed as a post-processing to ameliorate the obtained results. Initially, these results were obtained from a ML-based method. This system extracts the Semantic Relations, which are complicated or expressed through more than one word and it annotates them using a defined markup.

2.4 Evaluation matrices

To evaluate the performance of the system it is necessary to use well accepted performance measures such as precision, recall and F-measure. Precision is the proportion of number of correct relations extracted(N_{correct})to the total number of returned relations(N_{response}); recall is the proportion of correct relations extracted(N_{correct}) to the total number of relations which are corrected manually((N_{key})) ; and F-score is defined as the harmonic mean of precision and recall.

$$precision(p) = \frac{N_{\text{correct}}}{N_{\text{response}}} \quad \text{...}(2.1)$$

$$recall(r) = \frac{N_{\text{correct}}}{N_{\text{key}}} \quad \text{...}(2.2)$$

$$F - score = \frac{2 \times (precision \times recall)}{precision + recall} \quad \text{...}(2.3)$$

2.5 The Amharic language

Amharic is a Semitic language spoken predominantly in Ethiopia. It is the working language of the country. The language is spoken as a mother tongue by a large segment of the population in the northern and central regions of Ethiopia and as a second language by many others.Following the Constitution drafted in 1993, Ethiopia is divided into nine independent regions, each with its own regional language. Then, Amharic become the official or working language of several of the states regions within the federal system, including Amhara ,Gambella,Benshangul and the multi-ethnic Southern Nations, Nationalities and Peoples region. It is the second most spoken Semitic language in the world next to Arabic and the most commonly learned second language throughout Ethiopia. [42] It is the second largest language in Ethiopia (after Afan Oromo, a Cushitic language) and possibly one of the five largest languages on the African continent[9]. As a result it has official status and used nationwide. Despite it has large speaker population, the language has little computational linguistic resources.

2.5.1 The Amharic Writing

As it stated in [8]Amharic is written using a writing system called fidel - ፊደል ("alphabet", "letter", or "character") adapted from Ge'ez (the liturgical language of the Ethiopian Orthodox Church) language. In modern written Amharic, each syllable pattern comes in seven different forms (called orders), reflecting the seven vowel sounds.These seven orders (the first basic order and the other six orders) represent the different sounds of a consonant-vowel combination (a characterization known as syllabic).The non-basic forms are derived from the basic ones by somewhat regular modifications. the alphabet is written from left to right, in contrast to some other Semitic languages. There are 33 basic characters, each of which has seven forms called orders depending on which vowel is to be pronounced in the syllable. The seven orders were represent seven vowel sounds. Therefore, these 33 basic characters with their seven forms will give 7*33=231 syllable patterns (syllographs), or fidels. In addition to the 231 characters, there are other non-standard alphabets which contain special features usually representing labialization.

2.5.2 Amharic word categories

Based on the recent works(Baye,2000) as it is cited in [8] the Amharic language has five word categories based on the role of words in syntax, which means by considering the clear role of words in Amharic grammar. These five categories ስም (noun), ግስ (verb), ቅፅል (adjective), ተውሳከ ግስ (Adverb), and መስተዋድድ (preposition).

Noun: Like English, Amharic nouns are words used to name or identify any of a class of things, people, places, organization or ideas or a particular one of these.[3].A word will be categorized as a noun, if it can be pluralized by adding the suffix ኦችሯዎች and used as nominating something like person, animal, and so on [1].In Amharic sentences noun is used as to indicate subject of a sentence.Pronoun is a word that is used instead of a noun or noun phrase. They are characterized based on number, gender and possessiveness. Some of pronouns for deictic specifier such as ይህ እስዋ እኔ አንቺ እሱ እነሱ ..... ; Quantity specifiers such as አንዳንድ ጥቂት ብዙ .... and possession specifier such as የእሱ የእኔ የእነሱ ..... .

Verb: it is described by [8] as 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. As a result of this property a word at the end of such a sentence is expected to be tagged as a verb by an Amharic tagger.Verb expresses accomplishment of an action and used to close the sentence. For example in a sentence “ተሰማ ከባህርዳር መጣ” the word “መጣ” is verb since it appears at end of the sentence and closes the meaning of sentence.

Adjective: Adjectives in a sentence modify nouns to denote quality of a thing; that is, it specifies to what extent a thing is as distinct from something else,Adjectives in Amharic usually precede the nouns that they modify or describe to qualify a noun with some form of size, kind and behavior.For example in the sentence “ሰነፍ ተማሪ” (lazy student) the word “ሰነፍ ” (lazy)used to qualify the size of the noun ተማሪ(student). In this example, the adjective ሰነፍ “lazy” precedes the noun ተማሪ “student” which it modifies.

But [3] states that this does not mean that a word is an adjective just because it precedes a noun. For instance, in ይህ በግ “This sheep”, the word ይህ “This” precedes the noun በግ ‘Sheep’. Although the word ይህ functionally shares the feature of an adjective (modifier), it is a pronoun, a demonstrative pronoun.

Adverb: used to qualify a verb by adding extra idea on the sentence.In Amharic, adverbs can be found in either primitive forms (i.e. as separate words that appear by their own) or in compound forms as combinations of prepositions and some other words but that appear as separate or in rare cases as compound words. In each case they refer to place, time, circumstance etc[3]. Adverbs usually precede the verbs they modify or describe. The followings are some Amharic adverbs ትናንት ፣ ገና ፣ ዛሬ, አደገኛ ፣ ሃይለኛ ፣ተንኮለኛ.....

Preposition: Preposition is a word which can be placed before a noun and perform adverbial operations related to place, time, cause and so on; which can’t accept any suffix or prefix; and which is never used to create a new word. Prepositions have meaning only when they are attached or used together with other words such as nouns, verbs, pronouns and adjectives but they don’t have any meaning alone.For example in the sentence " አበበ በመኪና ወደ ቤት ሄደ” words በ ፣ወደ are prepositions. Some of prepositions include ከ ፣ ለ ፣ ወደ ፣ ስለ ፣ እንደ ፣ጋር.... prepositions can appear in different form as stated in [3]: some Simple prepositions are stand alone as separate words. for example ስለ ትምህርት ወደ ቤት and other Simple prepositions are prefixed or attached with other words (e.g. nouns and verbs). for example በመኪና ለሃገር

2.5.3 Amharic Punctuation marks and Numerals

In Amharic, there are different punctuation marks used for different purposes. In the old scripture, a colon (two dots ፡) has been used to separate two words. These days the two dots are replaced with white space. An end of a statement is marked with four dots (አራት ነ ጥብ ።) while ነጠላ ሰ ረዝ (፣ or ፥) is used to separate lists or ideas just like the comma in English and (ድርብ ሰረዝ ፤) is used as a semicolon in English. The question and exclamation marks have recently been included in Amharic writing system.

The Amharic number system consists of twenty single characters which are Geez numbers. They represent numbers one to ten, multiples of ten (twenty to ninety), hundred, and thousand. These characters are derived from Greek letters and in order to make them look like the Amharic characters the symbols are modified by adding a horizontal stroke above and below.
The system has no place value and there is not symbol representing the number zero (0). In addition, the number system does not use commas or decimal points. These situations make arithmetic computation using this system very complicated[8].
The paper presented in [3] also define Amharic Numerals as words representing numbers. They can be cardinal or ordinal numbers.In Amharic, the ordinal numbers are formed from the cardinal numbers by suffixing the suffix ኛ.Example

ሁለት two ሁለት- ኧኛ ሁለተኛ second

አስር ten አስር- ኧኛ አስረኛ tenth Like English, compound Amharic numerals are put separately. The following are examples to illustrate this.Example: ሁለት መቶ ሰላሳ አንድ ‘two hundred thirty one’

2.5.4 Problems in the Amharic Writing System

One of the problems in Amharic writing is the redundancy of symbols used with the same pronunciation.These different symbols give each word different meanings, in the Amharic language they have been used interchangeably.Table 1 which is presented in [8] shows an example of the character redundancy where more than one symbol is used for same sound.

Table 1: Amharic characters with the same sound

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

The other problem is related to the representation of compound Amharic words. Some compound words are represented as a single instance and other words are represented as two words including hyphen between them. For example if we take a word " ቤተ- መንግስት" (palace) this can be written as ቤተ መንግስት or ቤተመንግስት and this issue should be consider during the extraction process.

Chapter Three

3 Related Work

There have been different researches done on extracting relations between named entities for different languages. In this section we present some works done so far on extracting relations between named entities which are related with our work. Several methods have been proposed to extract relation between named entities. These methods can be broadly classified as rule based approach,machine learning based approach and hybrid approach.

3.1 Rule-based Approach

There are many research works done for different languages for extracting relations between named entities using syntactic and semantic patterns which is written manually by experts. The research work presented in [10] proposed a rule based approach for extracting relations between Arabic named entities. In this work the authors extracted a set of linguistic patterns from a training corpus that are rewritten in to local grammars implemented using Nooj platform. In this paper the authors extracted relations among five named entities (PERS PERS, PERS ORG, PERS LOC, ORG LOC, LOC LOC). To extract the relation between these named entities the authors define sub grammars containing the pattern of relation between each named entities. Which means they define a sub grammar for PERS PERS ,PERS ORG, PERS LOC, ORG LOC and LOC LOC relations.They defined four patterns of relation(sub grammars) (Family_ Relation, Social_ Relation, Business_ Relation and Communication) between PERS PERS named entities. The system used gender (masculine or feminine) and Number (singular or plural) features during the extraction of relations between named entities. The authors used 353 number of texts , 53197 number of word forms and 946 Arabic Named Entities for their test corpus. they evaluate their system for each entities pairs and they achieved a significant average result of 60%.

Another work done on [15] present the description and the experimental results of a novel rule based approach in mining the entities and its relations. The proposed method defines a new concept of entity relationship which treat entities relation as the relation of the main object and its supporting object. According to the authors the existence of supporting object is depend on the

existence of main object and there can be supporting object if there is a main object.
The relation between these objects are extracted through pattern learning process that utilize the Indonesia WordNet as an external knowledge. The goal of this study was to build a model that can be applied to the main object and supporting object relation extraction. In this research the authors did their experiment on various drug label documents which are collected from different web sites in order to extract drug name drug component. For relation extraction purpose the authors used pattern learning that is based on regular text expression which are found around targeted entities. The authors used Indonesian language WordNet as an external knowledge to generate a pattern from a training data that will be applied for testing data. The initial pattern constructed manually which contains relation words is used as initial input in addition to Indonesian language Wordnet.To extract entity relation the authors define an object relation which contains one main object and one or more of its supporting objects.In framework proposed by the authors there is no need to do the preprocessing or NLP tools that is commonly apply in many texts such as sentence parsing, stop word removal and POS tagging. The authors used 1566 total labeled drug documents in their experiment.

[46] Presented a novel rule based approach for extracting semantic relation between Spanish texts. In many relation extraction strategies a small variation in the punctuation and adjective modification would prevent the system from finding appropriate patterns. This can be solved by using a large corpus of training patterns or by applying parser that identifies the constitutes of the sentence as well as their syntactic function.But both of these solutions have their own problem. In the first case it is difficult and time consume task to prepare manually a large set of high quality training data and in the second case either parsers done for spanish performs a partial analysis or the parser is not freely avilable. so the authors in [46] proposed a partial parsing that simplifies the linguistic structures. The authors performed two experiments in order to know the performance of their relation extraction system based on real text. First the authors compared the rule based approach with a manually prepared corpus containing examples of the relation hasproffession. Second authors apply a parser with automatically obtained rules for hasprofession and hasbirthplace relation ship in the whole Spanish Wikipedia. In order to extract the sentence containing the related entities the authors obtain about 10,000 pairs for each relation from Spanish Wikipedia infoboxes.

Another work presented in [31] which is done on a medical domain also used a rule based approach for extracting semantic relation between medical entities with an empirical study on the treatment relation. In this work the authors designed an approach for medical entity recognition, relation construction and pattern extraction. These tasks are performed by two steps. First by extracting medical entities from sentence and determine their categories as for example disease, symptoms or treatment, and second, the authors extract semantic relations between extracted entities using lexical patterns. There is high terminological variation in the medical domain and this makes the recognition and categorization of medical named entities difficult.To solve this problem the authors proposed the solution in three steps. First they split biomedical texts into sentences and extract noun phrases by using LingPipe and Treetagger-chunker tools. Second Determine medical entities and semantic types with MetaMap. Finally they filter the obtained medical entities with a restriction on the semantic types used by MetaMap in order to keep only semantic types which are sources or targets for the targeted relations. To evaluate the system the authors used a total of 580 sentences which is collected from 20 medical articles. The evaluation result of precision and recall for the the extraction of treatment relations between a treatment (e.g. medication) and a problem (e.g. disease) is 75.72% and 80.4% respectively.

3.2 Machine Learning Approach

To fully automate the relation extraction system some researches have been done based on machine learning methods such as supervised, semi-supervised and un-supervised machine learning approaches.

3.2.1 Supervised Approach

This approach needs a fully labeled corpus and it considered relation extraction as a classification task.Classifiers are trained using a set of features selected after performing textual analysis of the labeled sentences. Depending on the nature of input to train the classifier supervised approach for relation extraction is further classified as feature based method and kernal method [27]. Feature based methods explicitly extract the lexical, syntactic and semantic features for statistical learning while kernel method does not explicitly extract features, rather it designs kernel functions over the structured sentences representations such as sequence, dependency or parse tree to capture the similarity between different relation instances.

[47] Propose a supervised learning method that uses contextual features based on centering theory as well as conventional syntactic and word base features for detecting a semantic relation between named entities.These features are organized as atree structure and used as an input for a boosting based classification algorithm.The authors used centering theory to determine how easily a noun phrase can be refered to in the following content and to detect relation between named entities.Unlike most previous researches [26, 34] this method extract the relation between named entities with in the same sentence or between entities found in different sentence.The authors used contextual feature in order to detect the semantic relations correctly when named Entity pairs located in a parallel sentence arise from a prediction ellipsis.Which means that the syntactic feature which is the path between two named entities in the dependency structure of the pair with a semantic relation is the same as with the feature of the pair with no semantic relation. As the authors described for instance the syntactic feature found in the sentence "Ken was born in Tokyo, Tom in New York." The feature of the pair with semantic relation "ken" and "Tokyo" is the same as the feature of the pair with no semantic relation "ken" and "New York".

The proposed method consists of the following three parts: (1) pre-processing like POS tagging, Named entity tagging and parsing (2) feature extraction which includes contextual, syntactic and word based features (3) classification.The authors used 1451 texts from Japanese newspapers and web blogs, whose semantic relations between person and location has been annotated manually by humans to test their proposed method. There were 5110 pairs with semantic relations out of 236,142 pairs in the annotated text. The experiments demonstrated that the proposed method has a precision of 73.7%, and recall 56.8%.

Another work presented in [44] used Supervised learning technique (SVM technique) to predict the word that can determine one or more semantic relations between two Arabic named entities. The main goal of the researchers was to detect a set of words that predicts relation between named entities. In this work there is no limit for the relations the system detects. An infinite number of relations without being dependent on predefined relations classes are detected.

In this work the authors extracted relations between four named entities(Person (PERS), Location (LOC) Organization (ORG) and Date (DATE)). The corpus used in this research contains 870 hetroginious articles gathered from various arabic electronic news papers such as Aljezira ,Al-Hyp and from wikipedia and 1,245 sentences containing at least a pair of NEs. For the sentence containing more than two named entities the authors treated each pairs of named entities separately.

[48] Presented a supervised approach using a classifier Support Vector machine to detect and classify the relation in Automatic Content Extraction (ACE) corpus.The classifier can be trained using a set of features including lexical tokens, syntactic structures, and semantic entity types selected after performing textual analysis (like POS tagging, dependency parsing, etc) of the labeled sentences.In addition to these linguistic features the authors used the distance between two entities to make the detection problem easier and to increase the performance of both the relation detection and classification. The authors divide the extraction task in to two sub tasks as relation detection and relation classification tasks. Relation detection is involved in identifying from every pair of named entities positive example of relations which can be fall into one of one of many relation categories such as Role, Part, At, Near and Social.
In relation classification the authors assigned a specific class to each detected relations. The authors proposed their relation extraction based on the following four assumptions:(1) Entities should be tagged beforehand so that all information regarding entities is available when relations are extracted. (2) Relations are binary, i.e., every relation takes exactly two primary arguments. (3) The two arguments of a relation, i.e., an entity pair, should explicitly occur within a sentence.(4) Evaluation is performed over five limited types of relations (Role, Part, At. Near, Social). The authors used 6140 for training and 1512 for development testing relation examples in ACE data set.The system is evaluated in terms of recall, precision, and F measure.

3.2.2 Semi Supervised Approach

The main problem with supervised methods are, they needs lots of tagged data for learning the classifier.If there is no enough annotated data to train and to many unannotated text for relation extraction then, the system will not give a good result. to solve these problems a semi supervised or boostrapping approaches for relation extraction have been gaining special attention.
These approaches require a very small amount of training data such as some seed instances which

are tagged manually. The system is trained with these seed instances and learn the classifier, and test with the classifier, and get more train examples by adding the test results to the training test.a sample of linguistic patterns or some target relation instances can be used to acquire more basic relations until discovering all the target relations.

[39] Is a semi supervised system developed by Zhou for semantic relation extraction between named entities using Labeled Propagation algorithm.Given a small amount of labeled data, the proposed method benefits much from the availability of large number of unlabeled data by first bootstraping a moderate number of weighted support vectors from all available labeled an unlabeled data using a co training procedure on top of support vector machine with feature projection.Then, a label propagation (LP) algorithm is applied to classify unseen instances by modeling the natural clustering inherit in both the labeled an unlabeled data via the bootstraped support vector and the remaining hard unlabeled instance after Support vector Machine bootstraping. Instead of propagating labels through all available labeled and unlabeled data, the proposed Label Propagation algorithm (LP) depend on weighted support vectors, bootstraped from all the available data, and the hard unlabeled instances, remaining in the unlabeled data after Support Vector Machine bootstapping which the author called critical instances.

In this paper, lexical, syntactic and semantic features multiple overlapping feature views which are called flat feature views are generated using random feature projection. For evaluation the authors used the ACE RDC 2003 which consists of 674 annotated text documents near to 300,000 words and 9683 instances of relations of training data and 97 documents near to 50,000 words and 1386 instances of relations of test data and ACE RDC 2004 corpora provided by LDC for evaluation which are collected from different news papers and broad casts.

Another work presented in[51] proposed a relation extraction system using semi supervised learning approach which is based on cluster based features. These clusters are selected by using several statistical methods.The main idea for this work is to extract relations based on word cluster. The assumption is that the absence of lexical features are compensated by the cluster feature. If the word (Named entity) have never been seen in the annotated relation instance (seed instance or training data), other words which share the same cluster membership with the word may have been observed in the relation instance.

This paper has two main contributions: (1) the authors explore the cluster based features in a systematic way and proposed several statistical methods for selecting effective clusters. (2) the authors study the size of training data on cluster features and analyze the performance improvements through an extensive experimental study. The authors focused on the extraction of relations between the basic seven ACE relation types: EMP-ORG, PHYS, GPE-AFF, PER-SOC, DISC,ART,OTHER-AFF.

The authors[51] employed a learning strategy which performs separately relation detection from relation extraction. For the training purpose first the authors trained a binary classifier to identify relation instances from non relation instances. Then rather than using the threshold out put of this binary classifier as a training data the authors used only the annotated relation instances a multi-class classifier for the 7 relation types which are defined by ACE. For testing the authors first applied the binary classifier to a given test instance for relation detection.If it is detected as a relation instance then they apply a multi-class classifier to classify it.

Another work presented in[50] proposed a semi-supervised technique that extracts binary relations between two Arabic named entities from the web using bootstrapping. This system depends on the pattern-based system. The system is evaluated in terms of precision and recall which is obtained from experiments made on four common relations these are: author-of (person, book) relation, president-of (person, country)relation, play-in (person, club) relation and CEO-of or chairman (person, company) relation . The relation type determine the value of precision and recall results but generally precision ranges from 61% to 75% and recall ranges from 71% to 83% but the best result is taken from a (player, club) relationship which gives 72% precision and 83% recall. The authors used Google search result summaries as an input to identify named entities and the relations between them. The proposed method is an iterative process in which each iteration consists pattern extraction and instance extraction phases. For pattern extraction first the authors select instance pairs (e1,e2) from the input instance then the proposed system searches for Arabic texts on the web with the selected seed instances in step one using Google search engine.For each input instance (e1,e2), download top 20 results that contains the two entities e1 and e2 in to a text file (candidate pattern sentence file). Then preprocessing on the candidate sentences is applied which includes normalization and sentence segmentation.Then to remove unrelated sentences sentence extraction validation is performed and the authors consider the sentence valid if it satisfy the following two conditions: (1) the sentence must contain both entities e1 and e2 in the same line (2) the maximum number of words between the two entities is not more than three.and finally patterns are discovered. The second phase is instance extraction phase and in this phase the system retrieves a set of newly extracted instances by receiving patterns extracted in the first phase.

3.2.3 Unsupervised Approach

Unsupervised method do not need fully annotated data or any initial manually selected seeds and are based almost entirely in clustering techniques and similarities between features or context words. According to [9] as they sited in [22] The goal of unsupervised learning is to group data in to clusters. in fact the basic task of unsupervised learning is to develop classification labels automatically. Unsupervised algorithms find out similarity between data in order to determine weather they can be characterized as a cluster.

[43] Proposed a tree similarity based unsupervised learning method to extract relations between named entities from a large raw corpus. The authors modified tree kernels on relation extraction to estimate the similarity between two parse trees efficiently using tree similarity function. Then the hierarchical clustering algorithm was used to group entity pairs in to different clusters based on their similarity. Finally each cluster is labeled by an indicative word as its relation type and unreliable clusters are pruned out. This work is based on the assumption that the same entity pairs in different sentence can have different relation types. This proposed method consists five steps: (1) Named Entity (NE) tagging and sentence parsing. In this case the authors used 150 hierarchical types and sub types of named entities from Sekine's named entity tagger and to generate a shallow parser they used Collin's parser. (2) Similarity calculation. for similarity calculation the authors used the minimum span parse tree including the named entity pairs when calculating the similarity function. (3) Named Entity pairs clustering. they used a bottom up hierarchical clustering method.Named entities are clustered based on the similarity score generated by tree similarity function. (4) Cluster labeling. they labeled each cluster by the most frequent word that contains the main meaning of a parse tree which is called head word. (5) Cluster pruning. Unreliable clusters are pruned out.

The authors used two criteria to identify which cluster is going to be pruned out. The first criteria is, if the relation type defined by the head word in the given cluster is not significant statistically. Second criteria is that, clusters whose Named Entity pair number is below a predefined threshold.

To test the proposed system the authors used The New York Times (1995) corpus. They evaluate their work on COMPANY-COMPANY (COM-COM) and PERSON-GPE (PER-GPE) named entity pairs.

Another work presented in [52] proposed an unsupervised approach to extract and cluster open relation between named entities from free text by re weighting word embedding and using the type of named entities as additional features. The proposed method consists of four stages: first they have performed Preprocessing. In this stage the authors extract named entities from each sentence in the data set using DBPedia spotlight and consider all sentences consisting atleast two named entities.

Second stage is feature extraction. In this stage features include word embeddings, dependency path between named entities and named entity types are extracted. The third stage is Spares Feature reduction and in this stage the authors used a Principal component analysis in order to avoid the sparse features.individual feature reduction of the sparse features is applied before merging them with the rest of the features. The final stage is Relation clustering. inorder to cluster the feature representation of each relation the authors used Hierarchical Agglomerative Clustering. Their evaluation result demonstrates that their method out performs than previous works in terms of f-score result.

Another work that uses unsupervised approach for relation extraction is presented in [54] which extracts a semantic relation between two named entities.The proposed models exploit entity type constraints within a relation as well as features on the dependency path between entity mentions to cluster equivalent textual expressions. Such that expressions in the same cluster bear the same relation type between named entities. They proposed a serious of generative probabilistic models which generate a corpus of observed triples of entity mention pairs and the surface syntactic dependency path between them. Their proposed models exploit entity type constraints within a relation as well as features on the dependency path between entity mentions. The authors evaluate their system in terms of measuring the clustering quality by mapping clusters to freebase relations.

3.3 Hybrid Approach

This is a mixed approach in which other approaches like rule based approach and machine learning approaches are combined. This approach can achieve an enhanced performance than either rule based or machine learning approaches achieve alone. Because of this recently research studies have been oriented to use this approach, for example [14, 55, 25].

The work presented in [14] proposed a hybrid method for extracting relation between Chines entities. These authors combined a supervised machine learning method with a rule based approach.They proposed a candidate sentence selecting method trained by conditional random field model for high-frequency relation words which can find out high quality training instances and reduce the mistakes produce by automatic tagging training data and improve the extraction performance.

To solve the problem of getting enough training data automatically for some rare relations the authors proposed a method based on some simple rules and knowledge base to extract these low frequency relation words. they construct their own Chines Semantic knowledge base by gathering a structured data which contains entity pairs from a chines encyclopedia. Then to extract a relationship they first traverse through their knowledge base to get the frequency of the relation word. based on the frequency number of the word they classified it as a high frequency relation word or as a low frequency relation word. If the word frequency number greater than 500 they classified it as a high frequency relation else it is classified as a low frequency relation. In the experiment a detailed comparison and analysis on some options of selecting candidate sentences are introduced.These candidate selecting method improves the average F1 value by 78.53%.

[55] Proposed a hybrid method to extract relations between diseases and treatments.these authors propose a method which combined a supervised machine learning based approach using Support Vector machine classification with a rule based approach. For the linguistic method a set of patters is constructed manually from the training corpus and from the MEDLINE corpora. In this set, a weight is associated with each pattern. This weight serves to choose the more convenient pattern in the case of multiple extraction candidates in the hybrid method. For the ML method, the authors investigated the SVM classifier, using lexical, morph-syntactic and semantic features. In addition to proposing a hybrid method for a medical domain that can efficiently extract relations between a diseases and treatments the authors also introduce a way to extract multiple relations from a single sentence.The presented approach takes advantage of the two techniques, relying more on manual patterns when few relation examples are available and more on feature values when a suf-

ficient number of examples are available. The obtained results of this hybrid approach show an enhancement toward the ML- and pattern-based methods.Their approach scores an overall 94.07% F-measure for the extraction of cure, prevent and side effect relations.

The work in[25] proposed a hybrid approach to relationship extraction from a stories.The proposed method combines Supervised learning method with a rules to extract relationships.The rules are used to the out put of the machine learning approach to improve the result.The method identifies the main characters and collects the sentences related to them.Then these sentences are analyzed and classified to extract relationships.

The proposed system has two phases. In phase one by using Naive Bayes Classifier The selected sentences for a pair of characters are classified in to corresponding relationship classes.Naive Bayes Classifier is trained by a CSV file, which contains a collection of sentences related to some common relationship.

In phase two the semantic similarity between selected sentences is measured with the sentence of CSV file. The general system architecture of the proposed system contains different processes including preprocessing of the input story. The proposed system removes these special characters as a preprocessing task. The text is then tokenized into words and also doing the PoS tagging. After preprocessing the next task is Named Entity Recognition.The proposed system is expected to identify three entity types namely person, location and organization using Stanford Named Entity Recognizer.Next to extracting named entities they performed anaphora resolution which is done using Stanford Deterministic Co reference Resolution System.Finally After doing named entity recognition and anaphora resolution, the whole story is preprocessed and segmented at sentence level. Sentences are selected if they include the specified tuple of characters in it.

The proposed system is evaluated using 100 short stories.from these a set of 300 character pair relations are taken. The system is evaluated in terms of precision and recall for parent-child and friend ship relations and it scored 86.6% precision and 81.6% recall for parent-child relation and 88.8% precision and 79% of recall for a friendship relationship.Generally the proposed system scored an average of 87% precision and average recall of 79.7%.

Chapter Four

4 System Design And Implementation

4.1 Introduction

In this section we discussed about the design and implementation of the proposed system, relationship extraction between Amharic named entities. The first part of this section is about our data set which is used for training and testing purpose in our system. The second section discus about the general over view of the proposed system architecture and finally detailed explanation of each modules along with their sub components are presented.

Generally our research method can be considered as part of a quantitative method. But specifically our research belongs to the sub category of quantitative research method which is experimental quantitative research method. In experimental research method the answer to research questions is obtained by conducting experiments.

4.1.1 Data Sets

Unlike other languages like English and Arabic Amharic is under resourced language so that it is difficult to find standardized annotated publicly available corpora. The lack of Amharic linguistic resources makes the machine learning technique, specially the supervised model difficult to use.

In our system we used two types of data, training data and test data. Training data is used for train the machine learning component. Then after the machine is trained and the model is created, the system has tested on test data to evaluate the performance of the system. As we have stated previously since there is no any publicly available annotated data for Amharic which is suitable for our system we drive to prepare our own data. We have constructed our corpus from different resources to obtain a suitable number of sentences. Our training corpus was gathered from different Amharic news sources such as, walta, fana, esat and zehabesha. Our corpus is composed of 764 sentences . The sentences that we collected has at least two named entities and we manually check it weather a sentence contains at least two named entities or not. This is because of two reasons.

The first reason is that our aim is to discover a binary relationship between named entities and another reason is if there is no any named entity in the training data, extracting the features and generating a good model is difficult and testing such kind of a model may degrade the performance of the system. Our data set is annotated with 10 relation types between named entities. These relations are: የትውልድ - ቦታ (Birth-Place), ጠቅላይ - ሚኒስቴር (Prime-minister), ሥራ- አስኪያጅ (Manager), ፕሬዝዳንት (President), የሚገኘው (Locted-in), መኖሪያ - ቦታ (Live-in), መስራች (Founder), ሃላፊ (Curator), ዋና- ከተማ (Capital-city) and መሪ (Leader). After preparing the corpus and performing pos tagging, dependency parsing and named entity tagging every word in a sentence kept in one line separated by tab space having features like pos tag, lemma, dependance and named entity type using the IOB notation.

Table 2: total data set

Relations (train,test)	Example
ፕሬዝዳንት 219(164,55)	ሩሲያ ፕሬዚዳንት ቭላድሚር ፑቲን
ጠቅላይ- ሚኒስቴር 122(91,31)	ዐብይ አህመድ ጠቅላይ ሚኒስትር ኢትዮጵያ
መስራች 76(57,19)	ቤተልሄም ጥላሁን መስራች ሶል ሪበልስ
ዋና- ከተማ 144(108,36)	ሉሳካ ዋና ከተማ ዛምቢያ
መሪ 20(11,9)	አቡበኸር አልባግዳዲ መሪ አይኤስ አይኤስ
ሊቀመንበር 36(19,17)	ኡብይ አህመድ ሊቀመንበር ኢጋድ
የትውልድ- ቦታ 17(11,6)	ስመኘው በቀለ የትውልድ ቦታ ማክሰኝት
የሚገኘው 90(68,22)	ባህር ዳር የሚገኘው ኢትዮጵያ
ሥራ- አስኪያጅ 18(11,7)	አረጋ ይርዳው ስራ አስኪያጅ ሚድሮክ
ሃላፊ 22(13,9)	ጃክ ዶርሴይ ሃላፊ ትዊተር
Total: 764(553,211)	

4.2 System Architecture

In this section the architecture of the proposed system will be presented. Figure 1 shows the over all architecture of Amharic named entity relation extraction system(ANERE). Our system contains three main components, preprocessing, training and testing. This three phases may be used in many information extraction especially relation extraction system and any other machine learning systems but the tasks which are performed in each phase are different depending on the data we used and the procedures in which the researches has chosen. Because of this, there is no any single architecture that we all should follow. Our proposed system architecture is presented as follow:

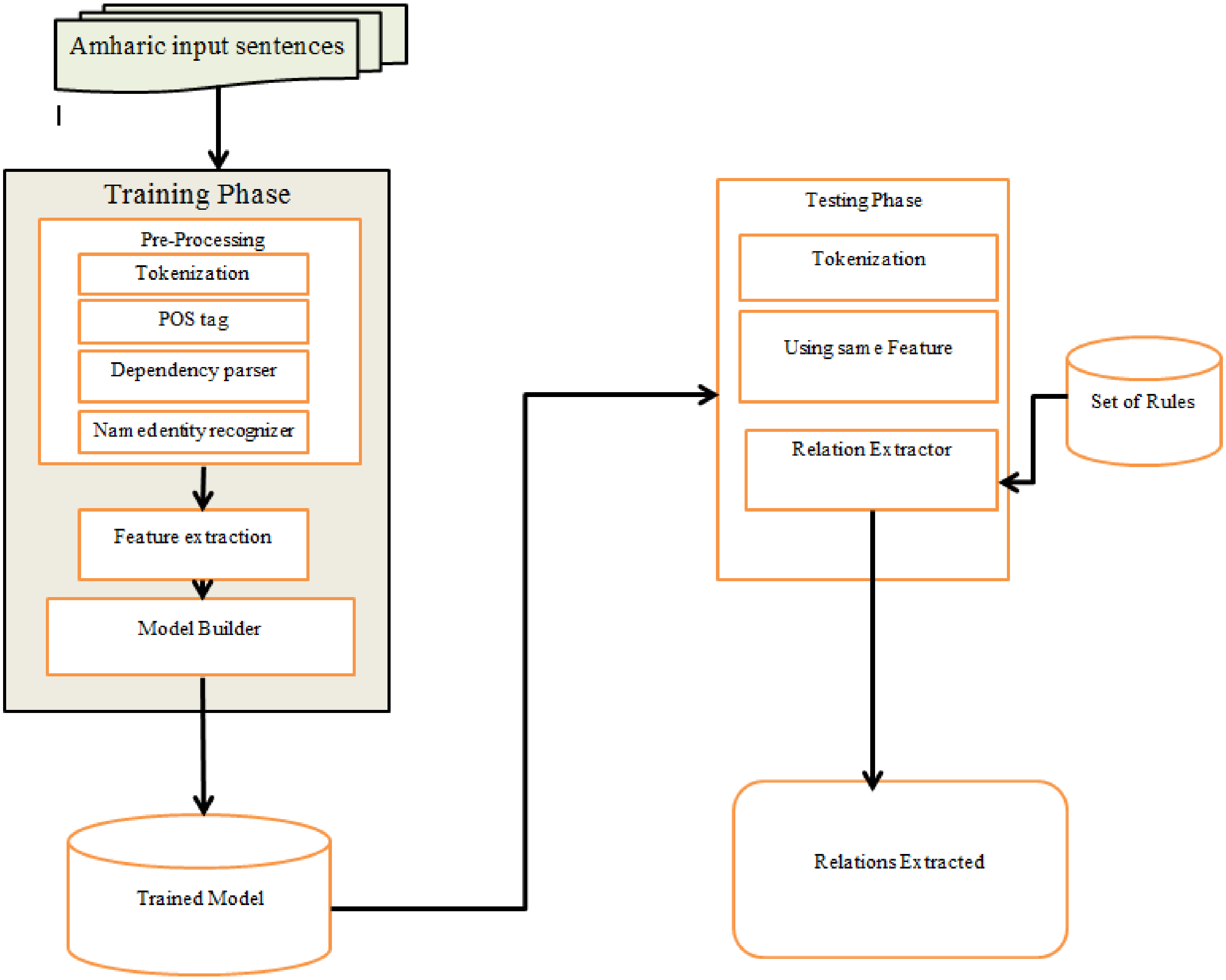

Figure 2: System architecture

4.2.1 Processes of ANERE

This Amharic named entity relation extraction system is designed in a manner that first it learns features and properties from the training data and then predicts possible relation types found between Amharic named entities by accepting Amharic plain texts. Our Amharic named entity relation extraction system architecture has pre-procesing, learning process and prediction processes.

4.2.1.1 Learning processes

Components in the learning processes are used to perform training. The training corpus is initially processed in our system. The corpus is tokenized and prepared in a tab separated format including different features like the word,word lemma, pos tag, parent word,dependency of words,named entity tag using bio format. These token are used by the feature extractor. The feature extractor extracts necessary features based on the tokens and tag sequences. These all extracted features are used then as an input for the model builder. After all the above tasks has been completed the model builder generate the trained model which used to predict relations found between Amharic named entities.

4.2.1.2 Preprocessing

4.2.1.2.1 Tokenization

Tokenization is the process of splitting a text in to its part of elements called tokens. It manipulates the text on the level of individual words. In Amharic a text can be split in to a tokens based on white space or other punctuation characters. Since in Amharic every word is separated by a white space it is simple to tokenize a given sentence. Tokenization is done on the training corpus during corpus preparation and also done on the input plain text during prediction process. The tokenization takes the input plain text from the user and split it in to a sequence of tokens that can make it easy for relation extraction.

4.2.1.2.2 POS tagging

The POS tagger receives input as a sequence of words of a sentence, and assigns POS tags to the words of that sentence. It identifies the grammatical form of a word based on the word itself and its surrounding context. Such grammatical forms output from the POS tagger are verb, adjective, noun, preposition etc.

To tag our data we used an Amharic tagger tool which is publicly available called "Habit Amharic tagger module" part of the HaBit project. The objectives of the HaBit project are to gather large scale text data(corpora) from the web for under resourced languages such as, Norwegian, partly Czech and major languages of Ethiopia -Amharic, Affan Oromo, Tigrinya, Somali and to build shallow processing applications.
The POS tagger for Amharic which is developed by the HaBit project was adopted a tree tagger and trained it by WIC corpus. We tested the system with many Amharic plain texts and it tagged them with a good accuracy.

4.4.1.2.3 Dependency Parser

The dependency parser is used to analyze the grammatical structure of a sentence. For our task since there is no any publicly available dependency parser for Amharic we used a parser which is done by Mizanu Zelalem as his research work. The parser performs with accuracy of 94.1%.

4.2.1.2.4 Named Entity Recognizer

The identification of named entities is the fundamental step in the extraction of relations between Amharic named entities. It can be accomplished by named entity recognizer. The recognizer tokenize the text and tags mainly three entities in the text, namely Person,Organization and Location. We used Stanford Named Entity Recognizer. It provides a default trained model and various models trained for different languages such as German, Spanish, French, Italian, Dutch and Portuguese for recognizing entities like Person, Organization and Location. In our case we trained it with our own data set.

4.2.1.3 Feature Extraction

In machine learning, feature extraction starts from an initial set of measured data and builds features intended to be informative and non redundant. Feature extraction is the process of transforming the input data in to a set of features which can very well represented in the input data. Features are properties of a text that are used to provide necessary information associated to a given word. Our feature extractor extracts all necessary features from the training data. This feature extractor provides necessary features for the process of building the model. The features we used are classified as entity features and word features. Entity features feature concerns about the entity type like pairs of entities in the sentence like PERSON-PERSON,PERSON-ORGANIZATION, PERSON-LOCATION,GPE-PERSON, PERSON-GPE, LOCATION-LOCATION etc. Word features are contextual features between named entities such as, the word of both the entities, words before and after named entities etc. Generally The features we used includes:

- Word or phrase of each sentence
- The word before and after each entity
- The words between the two entities
- POS tag of each words
- Dependency of a word
- Named entitiy tag of each words

All the extracted features are combined together and used as an input for the model builder which will predict the parameters of the model.

4.2.1.4 Model Builder

To build a trained model that will be used in prediction is the main concern of machine learning system. Based on the input from the training corpus including the extracted features which is generated from the feature extractor the model builder used to build a trained model. Generally a model builder is designed to generate a trained model that will be used in the prediction process.

4.2.1.5 Prediction phase

This is the final Phase in the Amharic named entity relation extraction system. It is the process of predicting relations from the given pre-processed token based on the trained model. The processed input text have the same format as the training data passed to the relation extractor. Then the relation extractor extracts relations found between an input text by the help of the trained model. Finally extracted relations are display as an output.

When extracting the relations, to compensate the error of the classifier we implemented a rule based approach. We first use the classifier to predict relationships between named entities, If we get the target tag that we want to predict we added in the annotation file, If we got the other tag which means the classifier did not succeed in placing the relation in one of the tags that it learned on,Then we go through the rules. some of the rules that we used includes :

- Check the left entity and the right entity values. For instance for ፕሬዚዳንት and ጠቅላይ- ሚኒስትር relation type the left entity should be a GPE and the right entity should be a PERSON.
- Check words before the first entity. For example if the first named entity is a GPE and if before it we have a word የቀድሞው then we apply the rule for ፕሬዚዳንት and ጠቅላይ- ሚኒስትር
- If between the named entities we see words like መዲና, ዋና ከተማ etc.Then we accept the words.

Chapter Five

5 Experiments

In this section,we have summarized the different experimental settings we used and present the obtained results with the classical measures of precision, recall and F-measure. In our case we have performed gold standard evaluation which is manually annotated. In this study the assumption is that this gold label will generally be correct. We manually annotated entities with their relations and compare the predicted result by our system with the gold standard annotated data.

5.1 Experimental Procedure

To evaluate the system manually annotated document has to be created. For our relation extraction system we prepared the annotated document with target relations. The general principle that we followed during evaluating the performance of the system is first we prepared manually annotated test data with a target relation types. Then the test data is given to the proposed system and the system would predict the relations found between named entities. Then we manually checks the out put generated by the system against the corresponding manual tags and compute the performance. We prepared test data with similar manner with the training data.

5.1.1 Corpus size

As we have described in 4.1.1 of this document our data both the training and test data is collected from different Amharic online news sources and totally 764 sentences are used. Our data set is annotated with 10 relation types between named entities. These relations are: የትውልድ - ቦታ (Birth-Place), ጠቅላይ - ሚኒስቴር (Prime-minister), ሥራ- አስኪያጅ (Manager), ፕሬዝዳንት (President), የሚገኘው (Locted-in), መኖሪያ - ቦታ (Live-in), መስራች (Founder), ሃላፊ (Curator), ዋና- ከተማ (Capital-city) and መሪ (Leader). Based on the number of annotated data that we collected for each relation types we realized that some relations like ሥራ- አስኪያጅ (Manager), ሃላፊ (Curator), የትውልድ - ቦታ (Birth-Place), መኖሪያ - ቦታ (Live-in) and መሪ (Leader) has small number of annotation examples and does not allow efficient learning driven extraction. Finally we choose only five relation types and the system is setup for the extraction of ጠቅላይ - ሚኒስቴር (Prime-minister), ዋና- ከተማ (Capital-city), ፕሬዝዳንት (President), የሚገኘው (Locted-in) and መስራች (Founder) relationships.

Table 3: initial corpus

Relations (train,test)	Example
ፕሬዝዳንት 219(164,55)	ሩሲያ ፕሬዚዳንት ቭላድሚር ፑቲን
ጠቅላይ- ሚኒስቴር 122(91,31)	ዐብይ አህመድ ጠቅላይ ሚኒስትር ኢትዮጵያ
መስራች 76(57,19)	ቤተልሄም ጥላሁን መስራች ሶል ሪበልስ
ዋና- ከተማ 144(108,36)	ሉሳካ ዋና ከተማ ዛምቢያ
መሪ 20(11,9)	አቡበኸር አልባግዳዲ መሪ አይኤስ አይኤስ
ሊቀመንበር 36(19,17)	ኣብይ አህመድ ሊቀመንበር ኢጋድ
የትውልድ- ቦታ 17(11,6)	ስመኘው በቀለ የትውልድ ቦታ ማክሰኝት
የሚገኘው 90(68,22)	ባህር ዳር የሚገኘው ኢትዮጵያ
ሥራ- አስኪያጅ 18(11,7)	አረጋ ይርዳው ስራ አስኪያጅ ሚድሮክ
ሃላፊ 22(13,9)	ጃክ ዶርሲይ ሃላፊ ትዊተር
Total: 764(553,211)	

Based on their frequency in the training corpus finally we have the following relation types that we are going to extract.

Table 4: Number of training and test sentences for selected relation types

Relations	Training corpus	Test corpus
ፕሬዝዳንት	164	55
ጠቅላይ- ሚኒስቴር	91	31
መስራች	57	19
ዋና- ከተማ	108	36
የሚገኘው	68	22

5.2 Performance evaluation

Many information extraction systems used standard evaluation metrics to evaluate the performance of their system. As described in chapter two of this document we used Precision, Recall and F-measure to evaluate the performance of our system.

Our evaluation is depend on correctly extracted relations, the total number of extracted relations by the system and manually annotated relations. Correctly extracted relations are those relations extracted by the system and are correctly matched with the manually annotated relations. Total number of extracted relations are those relations extracted by the system and which contains relation types which are not found in the manually annotated data and correctly extracted relation types. Manually annotated relations are relation types in which we want the system to extract and we used these data against the extracted relations by the system to evaluate the performance.

5.2.1 Results

The following table presents the results for each relation type. We have measured the performance using the test data.

Table 5: Precision P , Recall R and F-measure of each relation

Relationship type	Total No. relations	Total returned	Correct	Precision	Recall	F-measure
ፕሬዝዳንት	55	54	45	0.833	0.818	0.824
ጠቅላይ- ሚኒስቴር	31	39	30	0.769	0.967	0.856
መስራች	19	15	13	0.866	0.684	0.764
ዋና- ከተማ	36	35	33	0.942	0.916	0.928
የሚገኘው	22	19	15	0.789	0.681	0.73
Average				0.803	0.813	0.820

5.3 Discussion on the Experiment

In this paper, relationships are extracted from Amharic text by combining the supervised machine learning method with set of rules to improve the over all performance. The system is tested against a set of 163 sentences with 5 relation types such as ፕሬዝዳንት, ጠቅላይ- ሚኒስቴር, መስራች, ዋና- ከተማ and የሚገኘው. we tested the system separately with each relation types and we take the average result to measure the overall performance of the system. The experiment has done on standard evaluation metrics like precision, recall and f measure, and from our experiment we noticed that the highest precision is 94.2% for ዋና- ከተማ , the highest recall value achieved in this experiment is 96% for ጠቅላይ- ሚኒስቴር and the highest f score value is achieved 92% for ዋና- ከተማ. As it is shown in table 4 the proposed system gives an average precision of 80%, recall of 81% and an average f-score of 82%. The result showed that the proposed system performed well with the given input text.

Chapter Six

6 CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Currently the number of electronic data is increasing than ever before and we can find the high frequency of named entities in electronic texts. Because of large amount of electronic data it is difficult and tedious task to extract required information. To overcome the problem related to extracting and searching relevant and required information the concept of information extraction was introduced.

Named entity relation extraction is the process of extracting relations found between named entities. Many relation extraction systems have been developed for other languages like English, Arabic and any other foreign languages.To the best of our knowledge there is no any relation extraction system done for Amharic. Having a good relation extraction system can be used in question answering, text sumarization, semantic network, machine translation.

The purpose of this study is to develop a system for extracting relations found between Amharic named entities. The system contains different components like pre-processing component, training component, model builder component and finally testing component with respect to their sub components. In this study we presented a hybrid approach for extracting relations.The proposed approach relies on pattern based techniques and supervised machine learning technique. We implemented the machine learning component using an SVM classifier which uses different word features and entity features. To do this research a total of 764 sentences which are collected from different Amharic online news sources were used. These data set is classified to training and testing data set. We conducted our experiments on the test data which is prepared on the same way as the training data. We tested our system on five relation types. These is because of their frequency in the training data.The system achieved a promising value.

6.2 Contribution

The main contributions of the study are listed below:

- Model is designed using a hybrid approach for Amharic named entity relation extraction

- Developing Amharic named entity relation extraction system which can be used for many NLP and IE tasks

- We conducted experiments to demonstrate the performance and accuracy of the hybrid approach for Amharic named entity relation extraction

6.3 Recommendation

The task of relation extraction is very complex task for under resourced languages like Amharic. The developed Amharic named entity relation extraction system is not the last and best work in the area. It needs further improvements in order to develop a full fledged relation extraction system. The following are our recommendations they need to be taken in consideration for future works:

- In this work only relations found in sentence label are extracted. In fact relations can found over sentences and even across a document. Therefore in future extracting relations beyond sentence level by using co referencing may make the system more general.

- Our work is limited only extracting explicit relations. But as a future work we recommend that extracting also implicit relations make the system more acceptable.

- Due to time limitation and lack of available corpora we only focus on binary classification (extracting relations found between two named entities) and uni classification task. But to develop a full fledged system we recommend to extend it to multi classification task.

7 References

[1] ባየ ይማም ፣ “የአማርኛ ሰዋሰው” ፣ ሁለተኛ ዕትም ፣ አዲስ አበባ ፣ ጥቅምት2000 ISBN 978-99944-999-8-4.

[2] Dr Gunter Neumann, "Information Extraction Architecture And task Definition", DFKI and Saarland University

[3] A. A. Argaw, "AUTOMATIC SENTENCE PARSING FOR AMHARIC TEXT AN EXPERIMENT USING PROBABILISTIC CONTEXT FREE GRAMMARS," Msc, Computer Science, Addis Ababa University, Addis Ababa, 2002

[4] Takaaki Hasegawa, Satoshi Sekine, Ralph Grishman, "Discovering Relations among Named Entities from Large Corpora", Nippon Telegraph and Telephone(NTT) Corporation’s one-year visiting program , New York University, 2004

[5] Agnieszka Mykowiecka , Małgorzata Marciniak, Anna Kups c, "Rule-based information extraction from patients’ clinical data", Journal of Biomedical Informatics 42, 923–936, 2009

[6] Haiguang Li , Xindong Wu, Zhao Li, Gongqing Wu, "A relation extraction method of Chinese named entities based on location and semantic features", Springer Science+Business Media, 25 May 2012

[7] Boujelben I., Jamoussi S., et Hamadou A. B., “A hybrid method for extracting relations between Arabic named entities”, Journal of King Saud University-Computer and Information Sciences, 26(4). 425-440, 2014

[8] Bekele Worku Agajyelew, "Information Extraction from Amharic language Text: Knowledge-poor Approach", The School Of Graduate Studies Of Addis Ababa University In Partial Fulfillment For The Degree Of Masters Of Science In Computer Science, Addis Ababa, Ethiopia, 2015

[9] Besufikad Alemu, "A Named Entity Recognition for Amharic", The School Of Graduate Studies Of Addis Ababa University In Partial Fulfillment For The Degree Of Masters Of Science In Computer Science, Addis Ababa, Ethiopia, 2013

[10] BOUJELBEN INES, JAMMOUSI SALMA, BEN HAMADOU ABDELMAJID, "Rules based approach for semantic relations extraction between Arabic named entities", Conference: NooJ2012, At INALCO-Paris, Volume: Cambridge Scientific Publishers - CSP, 2012

[11] Atelach Alemu Argaw, Lars Asker, "An Amharic Stemmer : Reducing Words to their Citation Forms", Proceedings of the 5th Workshop on Important Unresolved Matters, pages 104–110, Prague, Czech Republic, June 2007

[12] Robert Dale, Kam Fai-Wong, Jian Su, "Natural Language Processing", Second International Joint Conference jeju Island, Korea, 2005 proceedings

[13] Jakub Piskorski and Roman Yangarber, " Information Extraction: Past, Present and Future", Multi-source, Multilingual Information Extraction and Summarization 11, Theory and Applications of Natural Language Processing, Springer-Verlag Berlin Heidelberg 2013

[14] Hao Wang, Zhenyu Qi, Hongwei Hao, and Bo Xu, "A Hybrid Method for Chinese Entity Relation Extraction", NLPCC 2014, CCIS 496, pp. 357–367, 2014

[15] Mujiono Sadikin, Ito Wasito, "A novel rule based approach for entity relations extraction", Journal of Theoretical and Applied Information Technology 20th April 2015 Vol.74 No.2

[16] Sunita Sarawagi, "Information Extraction", Foundations and Trends in Databases, Vol. 1, No. 3 (2007) 261–377, 2008

[17] Ralph Grishman, Beth Sundheim, "Message Understanding Conference - 6: A Brief History"

[18] Nancy A. Chinchor, "OVERVIEW OF MUC-7/MET-2", Science Applications International Corporation, San Diego, CA 92121

[19] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel, "The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation"

[20] Jing Jiang, "INFORMATION EXTRACTION FROM TEXT", Singapore Management University, Springer Science+Business Media, LLC 2012

[21] Beth M. Sundheim, "THE MESSAGE UNDERSTANDING CONFERENCES", Naval Command, Control, and Ocean Surveillance Center RDT and E Division (NRaD) Decision Support and AI Technology Branch San Diego, CA 92152-7420

[22] "Annotation Guidelines for Entity Detection and Tracking (EDT)", EDT-Guidelines-V4-2-6

[23] "Annotation Guidelines for Relation Detection and Characterization (RDC)", Version 4.3.2

[24] Rolf Sint, Sebastian Schaffert, Stephanie Stroka, and Roland Ferst, "Combining Unstructured, Fully Structured and Semi-Structured Information in Semantic Wikis", Siemens AG (Siemens IT Solutions and Services) Werner von Siemens-Platz 1 5020 Salzburg Austria

[25] Devisree V, P. C. Reghu Raj, "A hybrid approach to relationship extraction from stories", International Conference on Emerging Trends in Engineering, Science and Technology, ICETEST- 2015

[26] D. Zelenko, C. Aone, and A. Richardella, "Kernel Methods for Relation Extraction", Journal of Machine Learning Research, pages 3:1083–1106. 2003

[27] Mario Nuno Letria Ribeiro, "Extraction of non-taxonomic relations from texts to enrich a basic ontology", Thesis to obtain the master of Science Degree in Master Information Systems and Computer Engineering, June 2014

[28] Bowen Sun, "Named entity recognition Evaluation of Existing Systems", Master in Information Systems, Norwegian University of Science and Technology Department of Computer and Information Science, 2010

[29] Naincy Priya, Amanpreet Kaur, "A Name Entity Detection and Relation Extraction from Unstructured Data by N-gram Features", IOSR Journal of Computer Engineering (IOSR-JCE) Volume 17, Issue 4, Ver. II, July – Aug. 2015, PP 25-28

[30] Douglas E. Appelt and David J. Israel., "Introduction to Information Extraction Technology", Artificial Intelligence Center SRI International 333 Ravenswood Ave, Menlo Park, CA

[31] Asma Ben Abacha, Pierre Zweigenbaum, "Automatic extraction of semantic relations between medical entities: a rule based approach", Journal of Biomedical Semantics 2011

[32] Aron Culotta, Andrew McCallum, Jonathan Betz, " Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text"

[33] MUC, 1987-1998, "The nist MUC website: http://www.itl.nist.gov/iaui/894.02/relatedprojects/muc/"

[34] A. Culotta and J. Sorensen, "Dependency Tree Kernels for Relation Extraction", annual Meeting of Association of Computational Linguistics, pages 423–429, 2004

[35] Nguyen Bach, Sameer Badaskar, "A Review of Relation Extraction"

[36] Luke Zettlemoyer, "Relation Extraction", CSE 517 Winter 2013

[37] Xiao Ling, Sameer Singh, Daniel S. Weld, "Design Challenges for Entity Linking", University of Washington, Seattle WA

[38] Daniel Santos, Nuno Mamede, Jorge Baptista, "Extraction of Family Relations between Entities", L2F – Spoken Language Systems Laboratory – INESC ID Lisboa Rua Alves Redol 9, Lisboa, Portugal

[39] Zhou GuoDong , Qian LongHua, Zhu QiaoMing, "Label propagation via bootstrapped support vectors for semantic relation extraction between named entities", Computer Speech and Language 23, 464–478, 2009

[40] Natalia Konstantinova, "Review of Relation Extraction Methods: What Is New Out There?", University of Wolverhampton, Wolverhampton, UK AIST 2014, CCIS 436, pp. 15–28, 2014

[41] Fatma Ben Mesmiaa, Fatma Zidb, Kais Haddarb, Denis Maurelc, "ASRextractor: A Tool extracting Semantic Relations between Arabic Named Entities", 3rd International Conference on Arabic Computational Linguistics, ACLing 2017, 5-6 November 2017, Dubai, United Arab Emirates

[42] Alelgn Tefera, Yaregal Assabie, "Automatic Construction of Amharic Semantic Networks From Unstructured Text Using Amharic WordNet "

[43] Zhang M., Su, J., Wang, D.Zhou, G., Tan, and C.L., “Discovering Relations between Named Entities from a Large Raw Corpus Using Tree Similarity-Based Clustering”, In Proceeding of IJCNLP ’05, LNAI, vol. 3651, pp. 378–389, 2005

[44] Ines Boujelben, Salma Jamoussi, and Abdelmajid Ben Hamadou, " RelANE: Discovering Relations between Arabic Named Entities", pp. 233–239, 2014

[45] Alessandro Cucchiarelli and Paola Velardi, "Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence"

[46] Marcos Garcia and Pablo Gamallo, "A Weakly-Supervised Rule-Based Approach for Relation Extraction", Center for Research in Information Technologies (CITIUS), University of Santiago de Compostela

[47] Toru Hirano, Yoshihiro Matsuo, Genichiro Kikui, "Detecting Semantic Relations between Named Entities in Text Using Contextual Features", Proceedings of the ACL 2007 Demo and Poster Sessions, pages 157–160,

[48] Gumwon Hong, "Relation extraction using support vector machine", In Natural Language Processing–IJCNLP 2005, pages 366–377, Springer, 2005

[49] ZHOU GuoDong, SU Jian, ZHANG Jie, ZHANG Min, "Exploring Various Knowledge in Relation Extraction", Proceedings of the 43rd Annual Meeting of the ACL, pages 427–434, Ann Arbor, June 2005

[50] Shimaa M., Enas M, A.K. Al sammak and T.A. El-shishtawy, "Extracting Arabic Relations From The Web", International Journal of Computer Science & Information Technology (IJCSIT) Vol 8, No 1, February 2016

[51] Ang Sun Ralph, Grishman Satoshi Sekine, "Semi-supervised Relation Extraction with Large-scale Word Clustering", Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 521–529,2011

[52] Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, "Unsupervised Open Relation Extraction", arXiv:1801.07174v1 [cs.CL] 22 Jan 2018

[53] Bui, Q. C, "Relation extraction methods for biomedical literature",2012

[54] Limin Yao Aria Haghighi Sebastian Riedel Andrew McCallum, "Structured Relation Discovery using Generative Models", Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 1456–1466 Frederique Laforest1

[55] Ben Abacha, A. Zweigenbaum, P., "A hybrid approach for the extraction of semantic relations from MEDLINE abstracts." In: 12th International Conference on Intelligent Text Processing and Computational Linguistics CICLING2011, Tokyo, Japan, pp. 139–150,2011

[56] Jefferson Amado Pena, Victor Bucheli, Raul Ernesto Gutierrez de Pinerez, "Support Vector Machines for Semantic Relation Extraction in Spanish Language",13th Colombian Conference, CCC 2018, Cartagena, Colombia, September 26–28, 2018, Proceedings

[57] Wiem Lahbib, Ibrahim Bounhas, Bilel Elayeb, " A hybrid approach for Arabic semantic relation extraction", Association for the Advancement of Artificial Intelligence, 2013

[58] Rodrigo Agerri and German Rigau, "Robust multilingual named entity recognition with shallow semi-supervised features", Artificial Intelligence, 238:63–82. 2016

[59] Defense Advanced Research Projects Agency, 1995, Proceedings of the Sixth Message Understanding Conference (MUC-6), Morgan Kaufmann Publishers, Inc.

[60] Sudha Morwal, Nusrat Jahan and Deepti Chopra, "Named Entity Recognition using Hidden Markov Model (HMM)", International Journal on Natural Language Computing (IJNLC) Vol. 1, No.4, December 2012

[61] Yue Zhang and Jie Yang, "Chinese NER Using Lattice LSTM", Jul 2018

[62] Jianfeng Gao, Mu Li, Chang-Ning Huang and AndiWu, "Chinese word segmentation and named entity recognition", A pragmatic approach, Compu-tational Linguistics31(4):531–574, 2005

[63] Naji F. Mohammed and Nazlia Omar, "Arabic Named Entity Recognition Using Artificial Neural Network", Journal of Computer Science 8, 1285-1293, 2012

[64] National Institute of Standards and Technology, 2000, Automatic Content Extraction, http://www.nist.gov/speech/tests/ace/

[65] AbdelRahman S., M. Elarnaoty, M.Magdy and A. Fahmy, "Integrated machine learning techniques for Arabic named entity recognition", IJCSI Int. J. Comput. Sci., 7: 27-36, 2010

[66] Z.Kozareva, “Bootstrapping Named Entity Recognition with Automatically Generated Gazetteer Lists”, In Proceedings of EACL 2006, Trento, Italy, April 2006

[67] Lucia Specia and Enrico Motta, "A hybrid approach for relation extraction aimed to semantic annotations", Proceedings of the 2nd Workshop on Ontology Learning and Population, pages 57–64,Sydney, July 2006

[68] Boujelben, I., Jamoussi, S., Ben Hamadou, A., "Genetic algorithm for extracting relation between named entities", In: 6th Language and Technology Conference, LTC, Poznan ´ , Poland, pp. 484–488.

[69] Yaoyong Li, Kalina Bontcheva, Hamish Cunningham, "Adapting SVM for Natural Language Learning: A Case Study Involving Information Extraction", Department of Computer Science, The University of Sheffield, UK, 2009

[70] Aynadis Temesgen, "Design and Development of Amharic Grammar Checker",the School of Graduate Studies of Addis Ababa University in Partial Fulfilment for the Degree of Master of Science in Computer Science, Addis Ababa University, Ethiopia, 2013

[71] Kassa, Markos, “Implementing An Open Source Amharic Resource Grammar In Gf”, Chalmers University of Technology, Sweden, November 2010