Addis Ababa University

School of Graduate Studies

College of Natural Sciences

Department of Computer Science

Automatic Construction of Amharic Semantic Networks (ASNet)

By: Alelgn Tefera

Advisor: Yaregal Assabie (PhD)

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

March 2013

Addis Ababa University

School of Graduate Studies

College of Natural Sciences

Department of Computer Science

Automatic Construction of Amharic Semantic Networks

(ASNet)

By: Alelgn Tefera

Advisor: Yaregal Assabie (PhD)

Signature of the Board of Examiners for Approval

Name	Signature
1. Dr.Yaregal Assabie, Advisor	
2. Dr. Dida Midekso	
3.	

Dedicated to:

1. **Tefera Bogie (Father)**
2. **Maritu Muluneh (Mother)**

ACKNOWLEDGMENTS

Above all, I praise God, the almighty for providing me this opportunity and granting me the capability to proceed successfully. This thesis appears in its current form due to the assistance and guidance of several people. I would therefore like to offer my sincere thanks to all of them.

First and foremost I offer my sincerest gratitude to my advisor, Dr Yaregal Assabie, who has supported me throughout the thesis work with his patience and knowledge while allowing me the room to work in my own way by showing me the root to proceed. I attribute the level of my thesis accomplishment to his encouragement and effort and without him this thesis would not have been completed.

I would like to thank Mr. Solomon Getachew for his support on evaluation of the Amharic WordNet and test results concerning the linguistic parts.

I thank also Mr. Andargachew Mekonen for his support and preserving the stop words, he collected from different sources, used in this thesis.

Last, but by no means least, I thank my families and friends for their support and encouragement throughout my thesis work.

Table of Contents

Contents **Page**

FIGURES

TABLES

LISTINGS

ACRONYMS AND ABBREVIATIONS

FOPC:	First order predicate calculus
FTC:	Frequency Term cell
IR:	Information Retrieval
LSA:	Latent Semantic Analysis
NLP:	Natural Language Processing
POST:	Part of speech Tagger
PWMI:	Point wise mutual information
RDF:	Resource description framework
SVD:	Singular Value Decomposition
TF-IDF:	Term Frequency Inverse Document Frequency
VSM:	Vector Space Model
WSM:	WordSpace Model

ABSTRACT

Semantic networks are becoming popular issues these days. Even though this popularity is mostly related to the idea of semantic web, it is also related to the natural language applications. Semantic networks allow search engines to search not only for the key words given by the user but also for the related concepts, and show how this relation is made. Knowledge stored as semantic networks can be used by programs that generate text from structured data. Semantic networks are also used for document summarization by compressing the data semantically and document classification using the knowledge stored in it. As a result, semantic networks have become key components in many NLP applications.

In this thesis, we focused on the construction of semantic networks for Amharic text. We have developed Amharic WordNet as initial knowledge base for the system and extracted intervening word patterns between pairs of concepts in the WordNet for a specific relation from free text. For each pair of concepts which we know the relationship contained in Amharic WordNet, we search the corpus for some text snapshot between these concepts. The returned text snapshot is processed to extract all the patterns having n-gram words between the two concepts. We have used the WordSpace model for extraction of semantically related concepts. The process of relation identification in among these concepts utilizes the extracted text patterns. "Part-of" and "type-of" relations are very popular and frequently found between concepts in any corpus. We have designed our system to extract "part-of" and "type-of" relations between concepts.

The system was tested in three different phases with different datasets from Ethiopian News Agency and Walta Information Center. The accuracy of the system to extract pairs of concepts having "type-of" and "part-of" relations is 68.5% and 71.7% respectively.

Keywords: Amharic pattern extraction, Amharic relation extraction, Amharic Semantic Network, Amharic Knowledge base

CHAPTER ONE
INTRODUCTION

1.1 Background

Semantics is the study of meaning of a text. Its study mainly focuses on the relation between objects. Linguistic semantics is the study of meaning that is used by humans to express it using language [1, 2, 30, 35].

A semantic network is a network which represents semantic relations among concepts [1, 6]. It is often used as a form of knowledge representation. The nodes represent objects or concepts and the directed and labeled links represent relations between nodes. Thus, a semantic network is a directed graph. The nodes are usually represented by circles or boxes and the links are drawn as arrows between the circles. Concepts are the abstract representations of meaning of terms. In the phrase, "በ ኢፌዴሪ የ ትምርት ሚኒ ስ ቴር" which means "FDRE Ministry of Education", the term የ ትምርት ሚኒ ስ ቴር (Ministry of Education) shows a concept which represents an institution that establishes policies, strategies, rules and regulation of education in the country. The physical representations of concepts are terms. A term in turn can be represented by a word, phrase, sentence, paragraph, or document. So, we can get concepts at word, phrase, sentence, paragraph, or document level. We focused on extracting concepts at word level.

The main relations between concepts that are used in semantic networks are as follows [6]:

- synonym –two concepts A and B express the same thing
- antonym – concept A expresses the opposite of concept B
- meronym, holonym - part-of and has-part relations between concepts
- hyponym, hypernym - inclusion of semantic range between concepts in both directions
- Morphological relations- How the root word is related to its different inflections

Synonym

Two expressions are synonymous if the substitutions of one for the other never changes the truth value of a sentence in which the substitution is made [19]. By that definition, true synonyms are rare, if they exist at all. Two expressions are synonymous in a linguistic context *'C' if* the substitution of one for the other in 'C' does not alter the truth value.

Antonym

If two word forms are opposite to one another, then they are said to be antonyms [19]. Antonyms are a lexical relation between word forms. There is no semantic relation between antonyms. For example, the opposite of tall is short, so tall is the antonym of short. No semantic relation between them but relation between their forms exists.

Hyponym

Hyponymy/hypernymy is a semantic relation between word meanings [19]: for example, {Addis Ababa} is a hyponym of {Ethiopia}, and {Ethiopia} is a hyponym of {Africa}. In WordNet, much attention has been devoted to hyponymy/hypernymy (subset/superset, or the ISA relation). A concept represented by the synset {x, x′, . . .} is said to be a hyponym of the concept represented by the synset {y, y′, . . .} if the first concept is a "kind of" the second concept[6]. The relation can be represented by a pointer from hyponyms to its hypernymy.

Meronym

Another semantic relation is the "part-whole" (or hasA) relation known as meronym/holonym [19]. A concept represented by the synset {x, x′, . . .} is said to be a meronym of the concept represented by the synset {y, y′, . . .} if the first concept is a "part of" the second concept [6].

The above relations can be represented in semantic networks by pointers (labeled arcs) from one concept to another. The nodes in the network represent concepts and pointers represent relations. These relations represent associations that form a complex network; knowing where a term is situated in that network is an important part of knowing the term's meaning.

Morphological Relations

An important class of lexical relations is the morphological relations between word forms [6]. How the root word is related to its different inflections is part of morphological relation analysis. For example the words, ኢትዮጵያዊ, የኢትዮጵያ, ኢትዮጵያውያን are different inflections of the word ኢትዮጵያ (Ethiopia). The term ኢትዮጵያ (Ethiopia) represents a concept of a country found in Africa and the term ኢትዮጵያዊ (Ethiopian) shows a person who has an Ethiopian citizenship.

Implementation Standards

Semantic networks are designed and implemented based on the principles of resource description framework (RDF) [5] and other standards which state about the structure of the development of any ontology knowledge base. Computer implementations of semantic networks were first developed for artificial intelligence and machine translation, but earlier versions have long been used in philosophy, psychology, and linguistics [30].

Figure 1.1 shows an example of a semantic network for the Amharic sentence, አዲስ አበባ ከተማ የአፍሪካ ህብረት ዋና መቀመጫ ነች::

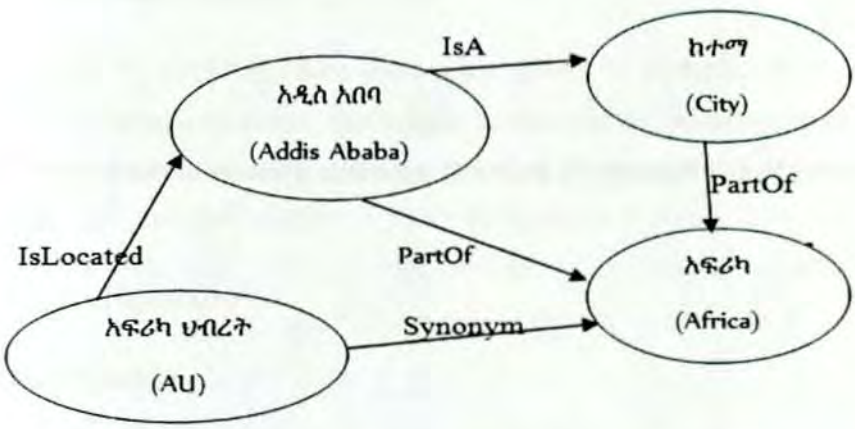


Figure 1.1. Example of Semantic Network

From the above graph, አፍሪካ ህብረት (Africa Union), አዲስ አበባ (Addis Ababa), ከተማ (City) and አፍሪካ (Africa) are noun concepts; and "ISA", "ISLocated", "PartOf", "Synonym" are relations between these noun concepts. For example, አዲስ አበባ(Addis Ababa) is a type of (ISA) a ከተማ(city) and አፍሪካ (Africa) has ከተማ(cities).

Semantic networks have different applications in different natural language processing activities. The following area of applications of semantic networks can be highlighted [21, 26].

- Semantic networks can be used in search engines for query expansion. Knowledge in the semantic network allows searching not only for the key words given by the user but also for the related concepts, and show how this relation is made.
- Semantic data can be used to describe documents (assigning tags) in NLP.
- For machine translation, semantic networks can provide a way to translate concepts easily.

- Semantic networks can be used for document clasification activities based on the knowledge it contains about the domain of each document.
- It also uses for document summerization by compressing the information semanticaly.

1.2 Statement of the Problem

Amharic language is one of the languages that has its own alphabets and syllabic patterns. Even though the language has such features, no sufficient research has been done so far to make use of possible natural language processing applications. Semantic networks have many benefits for machine translation, query optimization, document classification, language teaching & translation, and information retrieval.

Researches focusing on knowledge base construction should be attempted to have the above applications for the Amharic language. So the main intension of this research is to build a model for automatic construction of semantic networks, as a form of representation of knowledge base, for the language.

1.3 Objectives of the Study

1.3.1 General Objective

The general objective of this research work is to design a model for automatic construction of Amharic Semantic Network (ASNet) from free text corpus.

1.3.2 Specific Objectives

The specific objectives of this research are:

- To collect document corpus
- To build Amharic wordnet as a background knowledge for construction of ASNet
- To design and adopt algorithms
- To implement the model for prototyping purpose
- To evaluate the model

1.4 Methods

1.4.1 Litrature Review

Literature review has been done on different areas relevant to this research work. Approaches to automatic construction of semantic networks have been studied to unveil the current state of the art in the area. Based on literature, Amharic language has been studied with regard to its features that affect the automatic construction process of the semantic network. This includes the Amharic writing system and typical characteristics of the language.

1.4.2 Data Collection

We have used two different datasets in this research. The first which is obtained from Walta Information Center is composed of 1064 news items and all the news items are tagged with part of speech. This dataset is used for the extraction of concepts in the corpus used for the research. The second dataset was found from Ethiopian news agency and is composed of 3261 untagged free text news items. This dataset is used for the extraction of possible frequency of concepts that are extracted from the first tagged dataset.

1.4.3 Development Tools

The following tools were used in the experiment to develop the prototype of the model:

- Java programming language for implementation of algorithms.
- WordSpace model
- Manually built Amharic WordNet

1.4.4 Evaluation

There is no "gold standard" against which to evaluate the generated semantic networks. Ultimately, we believe that the best course of action is to use human evaluation, but once again the size of the network and the diversity of its representations cause problems.

It is not possible for an evaluator whether simply look at the network and determine if it is correct or read all of the inputs given to the system in order to determine if information is missing. But we tried to show, given concepts from Amharic WordNet, the efficiency of the

semantic analyzers (concept extraction based on semantic similarity) and network building algorithm (pattern based) of the model.

1.5 Scope and Limitations of the Study

Basically, semantic networks can be constructed using different semantic relational parameters. Synonym, antonym, type-of, part-of ... etc are some of the lexical relations between concepts in a knowledge base. There are also many relations between concepts in a certain domain. For example, we can get diseases and treatment relation in a medical domain.

In this research, the following points were incorporated:

- We have used single word terms as representation of concepts
- Amharic noun concepts were extracted from the corpus because other word categories like verbs and adjectives rarely occur as concepts. Verbs and adjectives mostly show the relation that exist betwean noun concepts.
- Synonym relation betwean concepts is constracted manually to increase accuracy of the construction of the semantic network and to increase the recall of the system. So, Amharic WordNet is built for such case as a background knowledge for our construction proccess.
- Part-of and type-of relations are the main and critical ones that exist between almost all concepts. They are important relations for the knowledge base construction. These relation were automatically costructed based on manually developed synsets (synonym concepts) of Amharic WordNet.

1.6 Applications of the Study

Semantic networks are basically a knowledge base for many applications in different areas [2, 26, 28]. ASNet can be used for different NLP and IR systems as a knowledge base component. It can be used for the following application areas of the language:

- Semantic search engine
- Document classification
- Language teaching
- Language translation- from Amharic to other local and international languages
- Word sense disambiguation

1.7 Thesis Organization

This paper is structured as follows. Chapter 2 presents literature review. Chapter 3 discusses different related works on automatic construction of semantic networks. Chapter 4 presents construction of Amharic semantic network. Test results were shown and discussed in Chapter 5. In Chapter 6, conclusions and future works were pointed out.

CHAPTER TWO
LITERATURE REVIEW

2.1 Semantic Networks

Semantic networks, as we said before in Chapter one, are used for knowledge representation as a form of graphs using concepts as nodes and semantic relations as labeled edges. It was developed in order to articulate Interlingua, a common language that would be used for translation between various natural languages [1, 6, 14]. Typical examples are WordNet and EuroWordNet [33] that describe relations between English words and defines the words using natural language. For example from Figure 2.1, a computer can have knowledge that "Fish is an animal" which "lives in a water". Here "Fish", "animal" and "Water" are concepts connected with each other by "are a/an" and "lives in" as relations. A semantic network involves three aspects [1]:

- Knowledge in which there are concepts and relationships among them.
- Diagrammatic representations like boxes, arrows and labels.
- a computer representation that allows database-like activity using algorithms that operate on these representations

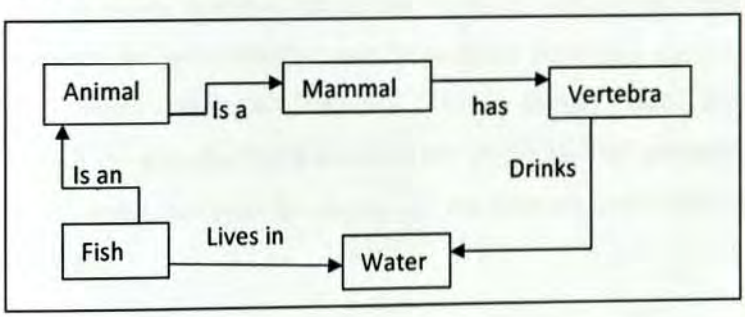


Figure 2.1. Example of Semantic Networks

2.2 Approaches to Automatic Construction of Semantic Networks

There are a number of researches done so far related to automatic construction of semantic networks. Those works implement different approaches to extract concepts and their relations automatically from free text and other resources. In this Section, we present knowledge based, corpus based and hybrid approaches [12].

2.2.1 Knowledge Based Approach

This is a method to extract relations between concepts in a supervised manner. Given a pair of target concepts, the technique is used to categorize and return the relation between the concepts using thesauri. Typically, using correct features of concepts as input to the system will produce a very good performance. It is a way of constructing knowledge bases like that of semantic networks based on the creation of semantic paths of thesaurus between words in a text [13, 19, 26, 32].

The extension of WordNet with semantic relations has added more promises of semantic network construction from text [6]. In this aspect, the method is very beneficial to use all of the available semantic relations in certain thesaurus. Figure 2.2 gives an example of the construction of a semantic network for two words "ti" and "tj" adapted from [26]. Assume the construction of a semantic path between senses "S.i.2" and "S.j.1" only (Initial Phase). Initially, the two sense nodes are expanded the semantic links available on WordNet. The semantic links of the senses, as found in the thesaurus, become the edges and the pointed senses the nodes of the network (Network Expansion).

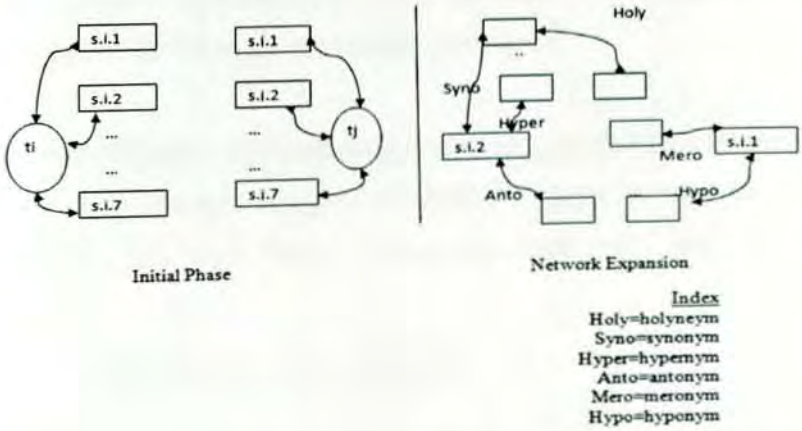


Figure 2.2. Constructing Semantic Networks from Word Thesauri

2.2.2 Corpus Based Approach

Knowledge based approach suffers from a number of problems. Manually prepared training data is very costly to produce and is limited in quantity. Because the relations are prepared on a particular corpus, the resulting classifiers tend to be biased toward that text domain [10, 26].

Semi supervised (Corpus-based) relation extraction leads to extract concepts and different relations between those concepts in large amounts of text. This approach can use very large amounts of data and extracts very large numbers of concepts and relations to construct semantic networks automatically. This approach is based on the distributional hypothesis, which states that similar terms tend to appear with similar contexts so that it is possible to group similar terms if their contexts are similar [10, 12, 32]. This approach can be considered as an open information extraction methodology for any text domain [16, 23].

Corpus based measures of word semantic similarity tries to identify the degree of similarity between words using information entirely derived from large corpora. There are two methods to do this: Pointwise Mutual Information [27] and Latent Semantic Analysis [13].

Pointwise Mutual Information (PMI)

PMI is a measure of relationship between two words or concepts in a corpus. It is a measure of how much one word tells us about the other word in a given data set. The value of the PMI can be positive or negative. Negative value means, the two words are independent- has no information content between them. It is used to find co-occurred words that occur together frequently in the corpus. If the value is positive, then there is a relation in between but, the strength of the relation depends on the magnitude of the value.

The PMI using data collected by information retrieval was suggested as an unsupervised measure for the evaluation of the semantic similarity of words. It is based on word co-occurrence using counts collected over very large corpora. Given two words "w1" and "w2", their PMI is measured as [27]:

$$PMI(w1, w2) = log_2 \frac{p(w1,w2)}{p(w1) \cdot p(w2)} \qquad (1)$$

P (w1, w2) is probability of (w1, w2) co-occurring in the corpus and p(x) is the probability of a word in the corpus where, 'x' stands for "w1" or "w2".

Latent Semantic Analysis (LSA)

LSA is a mathematical method that tries to bring out latent (hidden concept) relationships between words or phrases within a corpus. It focuses on co-occurrence of terms in a document and creating a term-document matrix. The key idea of LSA is to map terms (documents) to a vector space of reduced dimensionality.

The size of the term-by-document matrix is very large if the corpora are composed from many documents. The matrix may also be spares that may lead to some problem of mathematical computing. For dimensionality and spares data reduction, LSA uses singular value decomposition (SVD) algorithm on the term-by-document matrix 'T' representing the corpus. SVD [34] is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find co-relations among its rows and columns. So, here the semantic network will be constructed after the semantic analysis measures are performed by the methods discussed.

2.2.3 Hybrid Approach

This is the third approach which combines both the hierarchy of the used thesaurus, and statistical information for concepts measured in large corpora. In the corpus based approach, concepts and their relationships are often retrieved from their co-occurrence frequencies in text document collections. Machine readable dictionaries, lexicons and thesauri are some of the sources (pseudo-knowledge bases) having some rules inside. These manually built sources provide a natural framework for organizing concepts into some semantic classes [19, 26]. Using WordNet, many researchers have taken the advantage of this broad coverage lexical reference system to study concepts at word level and their relationships in between.

The integration of those manually built pseudo knowledge bases may complement the corpus based approach where "true" understanding of the text is very difficult to obtain. In such a way, the hybrid approach composed from the two approaches (knowledge based and corpus based) can take advantage of a labeled data structured manually as input and training data, while providing more statistical evidence to facilitate the distributional analysis process of any corpus [13, 26].

Using lexical information from manually built sources as input and extracts some lexico-syntactic patterns leads to extract other new instances of semantic classes. For a relation we

desire to extract, we can label some structured data and use this data as input to a system to extract the surrounding text patterns that show what feature (property) in a corpus, the input data has. So, this is the hybrid approach of the structured data (having rules inside) with that of free text corpus based approach which needs statistical algorithms to get additional computational evidence for the concepts and their relations.

Relation extraction systems usually use evidences (background knowledge) that are written explicitly in the input text to detect and characterize the semantic relations between target concepts. ASNet is constructed based on the hybrid approach by utilizing the available semantic relations from Amharic WordNet synsets and acquiring more statistical information from large corpus by using statistical algorithms.

2.3 Concepts Extraction Based on IR Models

Even though, there are other information retrieval (IR) models, we are interested to review the following two models which are used in this work.

2.3.1 Vector Space Model

Vector space model [27] is an algebraic model for representing text documents as a form of vectors. Documents and queries are represented as vectors. Suppose "d_j" , represents document vector for document 'j' and 'q' having words "w_1, w_2...w_t " represents query vector.

$$d_j = (w_{1,j} , w_{2,j} , \dots , w_{t,j}) \quad (2)$$

$$q = (w_{1,q}, w_{2,q}, \dots w_{t,q}) \quad (3)$$

Each dimension stands for a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing term weights have been developed. One of the best known schemes is term frequency-inverse document frequency (TF-IDF) weighting [27]. The definition of a term depends on the application. Typically, terms are single words, keywords, or longer phrases. If the words are chosen to be the terms, the dimensionality of the vector is the number of words in the vocabulary. Different vector operations can be used to compare documents with queries. Document similarity to the query is calculated using cosine of the angle between the document vector and query vector.

$$\cos\theta = \frac{d2.q}{||d2||\ ||q||} \quad (4)$$

Where $\mathbf{d_2}.\mathbf{q}$ is the dot product of the document "$\mathbf{d_2}$" and the query '**q**' vectors, $\|\mathbf{d_2}\|$ is the norm of vector "$\mathbf{d_2}$", and $\|\mathbf{q}\|$ is the norm of vector '**q**', $'\theta'$ is the angle between the vectors. The norm of a vector is calculated as:

$$||v|| = \sum_{i=1}^{n} vi^2 \quad \text{..... (5)}$$

The term weights in the document vectors are products of local and global parameters [27]. The weight vector for document '**d**' is:

$$V_{d=}[W_{1,d}, W_{2,d}, \ldots, W_{N,d}]^N \quad \text{..... (6)}$$

$$W_{t,d}=TF_{t,d}.\log\frac{|D|}{|\{d'\epsilon D|t\epsilon d\}|} \quad \text{..... (7)}$$

"$TF_{t,d}$" is term frequency of term 't' in document 'd' (a local parameter). "$\text{Log}\frac{|D|}{|\{d'\epsilon D|t\epsilon d\}|}$ " is inverse document frequency (a global parameter). "|D|" is number of documents in the corpus. "$|\{d'\epsilon D|t\epsilon d\}|$" is the number of documents containing the term, 't'. In such a way, the cosine of the angle between document "d_j" and query 'q' can be calculated as:

$$\text{sim(dj,q)}=\frac{dj.q}{||dj||\ ||q||}=\frac{\sum_{i=1}^{n} Wi,j \cdot Wi,q}{\sqrt{\sum_{i=1}^{n} W^2 i,j}\sqrt{\sum_{i=1}^{n} W^2 i,q}} \quad \text{..... (8)}$$

the value of "sim(dj,q)" is zero if the two vectors share no terms. The value is one if "dj" equals to 'q' in terms of contained terms.

2.3.2 WordSpace Model

WordSpace Model, concerns distributional semantics which is the practice of relating linguistic entities (like words, terms, phrases, sentences, documents) to each other based on their distributional properties in a corpora. It is completely data-driven, and requires no external resources [26].

The model is successful for many NLP tasks and applications, including vocabulary acquisition, information retrieval, machine translation, relation extraction, just to name a few examples [26, 32].

Using this model, semantic relations between concepts of words, are captured by collecting information on which words co-occur with similar other words. The basic input to the model is the co-occurrence frequencies of words in corpora. So, here many documents are needed to capture enough count of frequency of each word for better relation extraction.

The model was designed based on the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts [26]. According to this hypothesis, if we observe two words that constantly occur with the same contexts, we are reasonable in assuming that they mean similar things. This implies that it is not only words co-occur with each other but also possible that the words co-occur with the same other words.

In general, the WordSpace model is a spatial representation of word meaning. Its central idea is that semantic similarity can be represented as proximity in n-dimensional vector space matrix, where 'n' can be any integer ranging from 1 to some very large number which is determined by the number of columns in the matrix.

Even though, visualizing such high-dimensional spaces is sometimes difficult, we can get an idea of what a spatial representation of semantic similarity might look like if we consider a 1-dimensional or a 2-dimensional WordSpace, like the one represented in figure 2.3. In such a geometric interpretation, spatial proximity (nearness) between words indicates how they are related. For instance, both WordSpaces in figure 2.3 show 'ፖለቲካ' being closer to 'መንግስት' than to 'ኢትዮጵያ', which can be interpreted as a representation of meaning similarities between these words. In this case, the meaning of 'ፖለቲካ' is more similar to the meaning of 'መንግስት' than to the meaning of 'ኢትዮጵያ'.

The core idea behind WordSpace models is that words and concepts are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space. Such models are designed to represent words and documents in terms of underlying concepts (contexts).

In the WordSpace, the high-dimensional vector space is produced by collecting the data in a co-occurrence matrix 'F', such that each row "FT" represents a unique term 'T' and each column "FC" represents a context 'C', typically a multi-word segment such as a document, or another term. In the former case, where the columns represent documents, we call the matrix a term-document matrix, and in the latter case where the columns represent terms, we call it a term-term matrix. LSA and random projection algorithms like SVD, which are based on co-occurrences analysis, are used in WordSpace model.

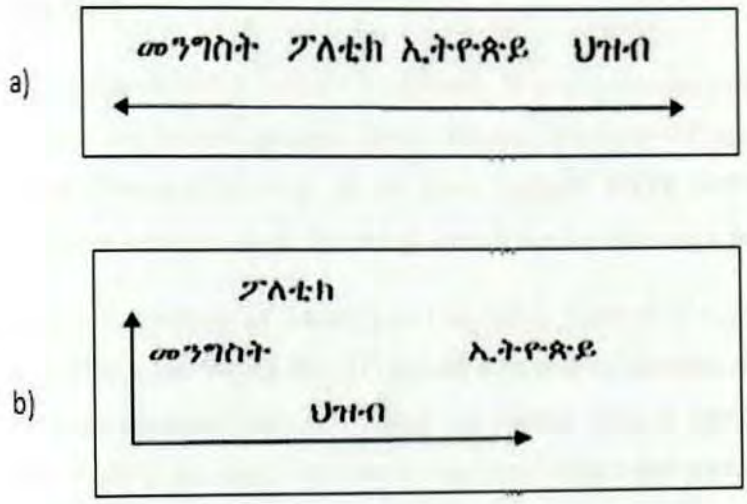


Figure 2.3. (a) 1-Dimensional, and (b) 2-Dimensional WordSpaces

The cells "FTC" of the co-occurrence matrix record the frequency of co-occurrence of term 'T' and document or term 'C'. As an example, if we use term-based co-occurrences, and observe a given term three times co-occurred with another term in the data, we enter 3 in the corresponding cell in the co-occurrence matrix.

The frequencies are usually normalized and weighted in order to reduce the effects of high frequency words. The point of the co-occurrence matrix is that the rows "FT" effectively constitute vectors in a high-dimensional space, such that the elements of the vectors are (normalized) frequency counts, and the dimensionality of the space is determined by the number of columns in the matrix, which is identical to the number of contexts (words or documents) in the corpora.

We call the vectors context vectors since the vectors represent the contexts in which words have occurred. Here now, the context vectors are representations of the distributional profiles of words in the data set, which mean that it is possible to define distributional similarity between words in terms of vector similarity.

Based on the distributional hypothesis discussed earlier, this makes it simple to compute semantic similarity between terms. It is possible to compare their context vectors using any of many possible vector similarity measures, such as the cosine of the angles between them.

2.4 Pattern Extraction

Text patterns [11] are appearance of n-gram [20] words in a sentence or passage with some frequency. For example the Amharic phrase, "በክልሉ (Region) ከሚገኙ ከተሞች (cities) መካከል ሐዋሳ (Hawassa)", to mean "Hawassa is a city in the given region". ከሚገኙ ከተሞች መካከል is the intervening words pattern between words, በክልሉ and ሐዋሳ. It is a 3-gram words pattern.

Text pattern extraction is a process of detecting and retrieving patterns of n-gram words using seed instances for specific need. Works like [24] present a method to automatically learn surface text patterns expressing relations between instances of classes using a search engine. Their method, based on a training set, identifies natural language surface text patterns that express some relation between two instances.

For instance, "...መካከል... (such as)" is a good pattern expressing the relation between instances ጋምቤላ(Gambela) of class "ክልል (Region)" and "ኢትዮጵያ (Ethiopia)" of class "አገር (Country)".

The way we extract text patterns depends on specific need. The result of the extracted patterns also depends on the input seeds. If we are interested to extract patterns that show organizations and their locations, we need to care more on selection of seed pairs, as input to the system, that are instances of the classes, organization and location. For example, to get text patterns that are used to extract organizations and their location in Ethiopia, we may use pair of instances, {AAU, A.A} as one input.

CHAPTER THREE
RELATED WORKS

3.1 WordNet: An On-line Lexical Database

WordNet is a simple semantic network constructed based on the lexical relations of English words and is motivated by current psycholinguistic theories of human lexical memory [19, 21]. In 1985 a group of linguists at Princeton University carried out to develop the WordNet. It was supposed to be a dictionary based on psycholinguistic principles. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying concept. Synonym sets can be linked by different lexical relations. The main relations in WordNet are; part-whole, hierarchical, antonym, synonym and morphological relations. WordNet divides the lexicon into five categories: nouns, verbs, adjectives, adverbs, and function words. This is the main difference with the normal alphabetical dictionary. WordNet contains only nouns, verbs, adjectives, and adverbs in its database. Table 3.1 shows the main relations and example of them in WordNet adapted from [19].

Table 3.1. Semantic Relations in WordNet

Semantic Relation	Syntactic Category	Examples
Synonym (Similar)	N, V, Aj, Av	(pipe, tube), (rise, ascend), (sad, unhappy), (rapidly, speedily)
Antonym (Opposite)	Aj, Av, (N, V)	(wet, dry), (powerful, powerless), (friendly, unfriendly), (rapidly, slowly)
Hyponym (Subordinate)	N	(tree, plant)
Meronym (Part)	N	(ship, fleet)
***N** = Nouns **Aj** = Adjectives **V** = Verbs **Av** = Adverbs*		

WordNet looks like a thesaurus because it is organized by semantic relations and attempts to organize lexical information in terms of word meanings, rather than word forms. Since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets. For example, if there is a semantic relation 'R' between meaning {x, x′ . . .} and meaning {y, y′ . . .}, then there is also a relation R′ between {y, y′ . . .} and {x, x′ . . .}.

3.2 MindNet: Acquiring and Structuring Semantic Information from Text

MindNet is a Microsoft product which is a lexical knowledge base constructed automatically from the definitions and example sentences into machine-readable dictionaries (MRDs) [25]. MindNet is a system that presents a general method for acquiring, structuring, accessing, and exploring semantic information from natural language text. It is developed in a fully automatic manner, based on the use of a natural language parser [25]. The extraction of the semantic information contained in MindNet exploits the parser used in the Microsoft Word 97 grammar checker. The parser produces syntactic parse trees and deeper logical forms, to which rules are applied that generate corresponding structures of semantic relations. The different types of labeled semantic relations in MindNet are given in the table 3.1.

Table 3.2. Current set of Semantic Relation Types in MindNet

Attribute	Goal	Possessor
Cause	Hypernym	Purpose
Co-Agent	Location	Size
Color	Manner	Source
Deep_Object	Material	Subclass
Deep_Subject	Means	Synonym
Domain	Modifier	Time
Equivalent	Part	User

The automatic extraction of semantic relations from sentence for MindNet produces a hierarchical structure of concepts along with their relations, representing the entire definition from which they came. Figure 3.1 shows semantic relation for the definition of "**car**" [25].

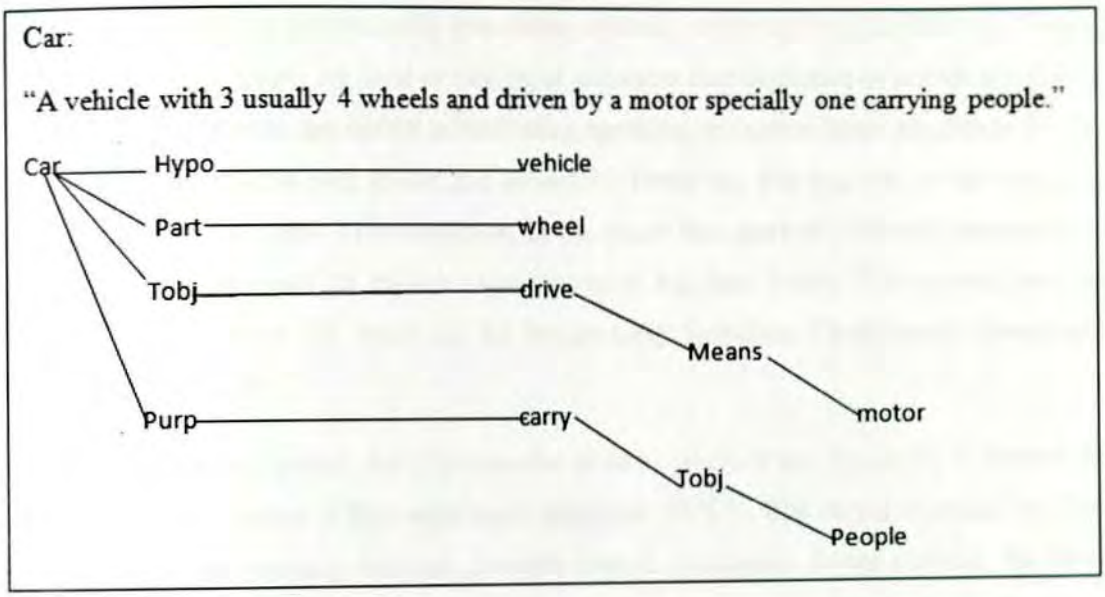


Figure 3.1. Semantic Relation Structure for a Definition of Car

3.3 Automatic Extraction of Semantic Networks from Text Using Leximancer

Leximancer is a system for processing of semantic relations of text data in language independent manner. The system provides unsupervised and supervised analysis using seeded concept classifiers [28] on free text corpus.

The strategy that the authors used in their work for conceptual mapping of text involves grouping families of words to thesaurus concepts. These concepts are then used to classify text at a level of several sentences. The resulting concept tags are indexed to present a document exploration environment for the user. A smaller number of simple concepts can index many complex relationships by recording co-occurrences. To achieve this, they developed some algorithms: a learning optimizer for automatically selecting, learning, and adapting a concept from the word usage within the text, and an asymmetric scaling process for generating a cluster map of concepts based on co-occurrence in the text.

3.4 ASKNet: Automated Semantic Knowledge Network

ASKNet is a system for automatically generating semantic networks from English text. Parsers and semantic analyzers [4] are used to turn input sentences into fragments of semantic network. These network fragments are further jointed using spreading activation-based algorithms [4, 36] and this algorithm utilizes both lexical and semantic information. The emphasis of the system is on wide-coverage and speed of construction. In this paper they showed a network consisting of over 1.5 million nodes and 3.5 million edges created in less than 3 days. Their system uses the Clark and Curran parser [7], based on the linguistically formalism Combinatory Categorical Grammar (CCG) [3, 7].

Once the data has been parsed, ASKNet uses the semantic analysis tool Boxer [3] to convert the parsed output into a series of first order logic predicates (FOLP). The output of Boxer is a first order predicates representing relations between objects (concepts). Boxer contains the basic semantic classes in a sentence to develop labeled and directed relations. As stated in the paper, ASKNet can efficiently translate Boxer's semantic output for each sentence into one or more semantic network fragments. The semantic networks created by ASKNet consists of object nodes linked by directed labeled relations.

The idea behind the ASKNet network is like the human brain, that concepts and relations need to be semantically linked so that thinking about (or firing) one concept leading other related concepts making them more likely to fire in the near future. They have used the spreading activation principle that, *by firing one or more nodes and analyzing the way in which activation spreads through the network; they can determine the semantic distance between various entities and concepts* [4]. This allows them to determine how closely related two entities or concepts are even if they are not directly linked. For example, the sentence "John saw Bob talk to Alice yesterday. Alice met Susan twice. Susan knows that Bob likes Fred." represented as semantic network [4].

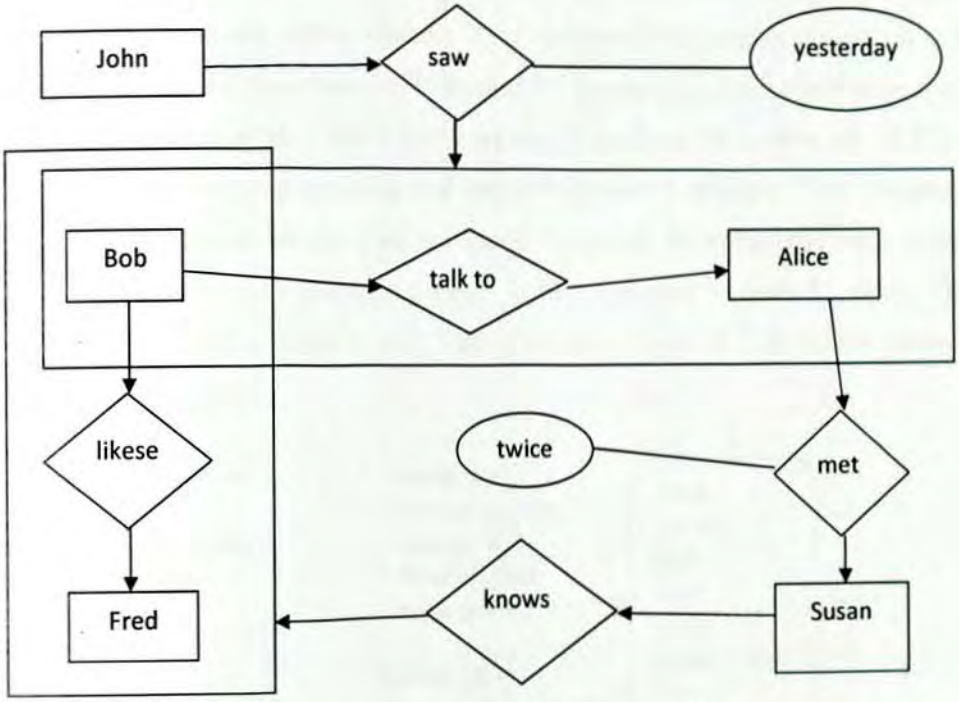


Figure 3.2. A simplified ASKNet Semantic Network

3.5 Extracting Semantic Networks from Text Via Relational Clustering

This is a system developed based on an unsupervised approach to extract semantic networks from large volumes of free text. They use the TextRunner [16, 23] system to extract tuples from text, and then fetch general concepts and relations from them by jointly clustering the objects and relational strings in the tuples. They have done experiments on a dataset of two million tuples and stated that it out performs three other relational clustering approaches [29], and extracts meaningful semantic networks. They present SNE (Semantic Network Extractor), a scalable, unsupervised, and domain-independent system that simultaneously extracts high-level relations and concepts, and learns a semantic network from text. It first uses TextRunner to extract ground facts as triples from text, and then extract knowledge from the triples.

Semantic Network Extraction

Semantic Network Extractor (SNE) simultaneously clusters objects (concepts) and relations in an unsupervised manner, without requiring the number of clusters to be specified. The object clusters and relation clusters respectively form the nodes and links of a semantic network. A link

exists between two nodes if and only if a true ground fact can be formed from the symbols in the corresponding relation and object clusters. They compared the various models on a dataset of about 2.1 million triples extracted in a Web crawl by TextRunner. Each triple takes the form r(x, y) where r is a relation symbol and x and y are object symbols. They have got 15,872 distinct r symbols, 700,781 distinct x symbols, and 665,378 distinct y symbols. Two characteristics of Text Runner's extractions are that they are sparse and noisy. To reduce the noise in the dataset, their search algorithm only considered symbols that appeared at least 25 times. This leaves 10,214 r symbols, 8942 x symbols, and 7995 y symbols. Figure 3.3 shows the learned concept and relation clusters [29].

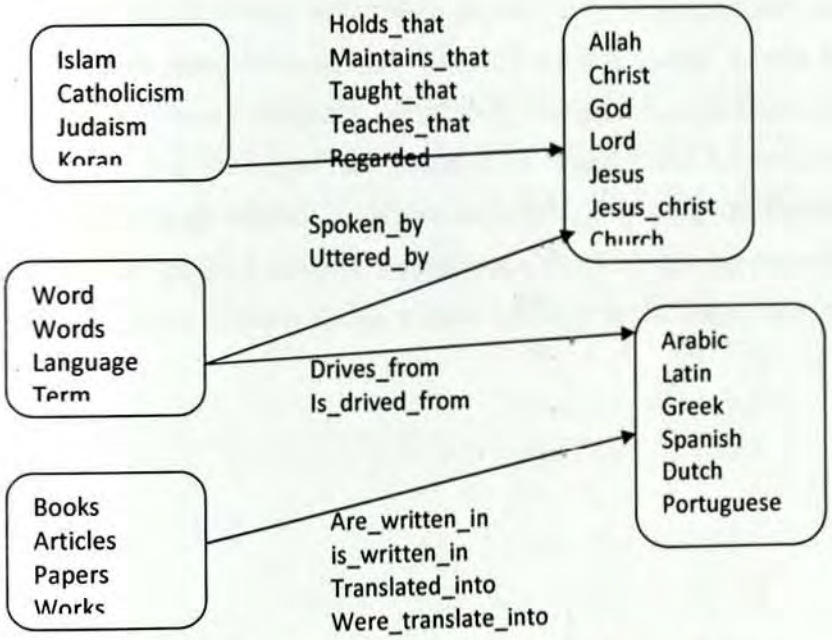


Figure 3.3. Fragments of a Semantic Network Learned by SNE

3.6 Automatic Thesaurus Construction for Amharic Text Retrieval

Andargachew Mekonen [18] attempted to develop Amharic thesaurus automatically from free text based on the WordSpace model paradigm. He used the Lucene indexer to index all the Amharic bible documents downloaded from web and WordSpace model to represent each index term as a form of vector fit for further geometric computations. He had developed thesaurus terms by searching the model for a given seed term using cosine function as a semantic measuring metric between term vectors. He had tested his system and got 58% accuracy using the Amharic Bible corpus.

Summary

In general, these works have been done on words as concepts using many NLP tools for English language. As English is a resource full language, we cannot compare and contrast those works with our work in advance. For example, ASKNet used English parser to extract different word categories (part of speech). It also used semantic analysis tool for extracting semantic relations between concepts. For Amharic language, no such resource is available to implement such systems easily. We tried to show the methodologies they followed and how we can apply these methodologies for Amharic language. Most of the papers show that extracting concepts statically from free text corpora is mandatory. Detecting and identifying the different relations between those extracted concepts is another independent process. In those papers, the relation extraction process has been done using different NLP tools of English language. Among these tools are: POS taggers, parsers, semantic analyzers, named entity recognizers, noun phrase extractors…etc. what we have used is only the tagged data instead of the tagger. Using the manually tagged data, we have used pattern based relation extraction methodology by using the manually developed Amharic WordNet for specified relations. So the reason why we presented these related works is to show that our approach is nearly similar to them and easy for languages like Amharic having limited NLP resources.

CHAPTER FOUR
CONSTRUCTION OF AMHARIC SEMANTIC NETWORK

4.1 Amharic Language

Even though many languages are spoken in Ethiopia, Amharic is dominant and is spoken as a mother tongue by a large segment of the population and it is the most commonly learned second language throughout the country [17]. The language is the working language of the federal government of the country.

4.1.1 Amharic Writing System

Three writing systems are in use in Ethiopia, the Amharic syllable, the Roman alphabet, and Arabic script [17, 31]. Amharic syllable is used for Ge'ez, Amharic, and Tigrigna, with slight modification. The Amharic syllable is uniquely Ethiopian writing system [31]. The writing system has a similarity with some Semitic languages like Arabic in having vowel marks added to basically consonant letters. The present writing system of Amharic is taken from Ge'ez. Ge'ez in turn took its script from the ancient Arabian language mainly confirmed in inscriptions of the Sabean dialect [17].

4.1.2 Characteristics of Amharic Language

Amharic language has some problems that do not enable us to do the information extraction process smoothly [17]. The first problem is the presence of "unnecessary" alphabets (fidels) in the language's writing system. These fidels have the same pronunciation but different symbols. These different fidels can be used interchangeably without meaning change. The fidels are ኧ and ዐ, ጸ and ፀ, ሰ and ሠ and ሀ and ኀ, ሐ, and ኻ. For example, the word "man" can be written as, ሰው, ሠው both mean the same, although they are written differently.

The other problem is in the formation of multi word concepts. Multi word concepts (compound words) are sometimes written as two separate words and sometimes as a single word. For example, the word "kitchen" can be written as "ወጥቤት" or "ወጥ ቤት". Unless they are normalized, the above words will be considered as to represent two different meanings. There are many such compound words, which need some effort to have a standard way of forming them.

Amharic is a morphologically rich language where up to 120 words can be conflated to a single stem [17, 18]. The word units of Amharic are phoneme, morpheme, root, stem, and word. The 34 base characters are phoneme. A collection of phonemes forms morphemes, which is the smallest meaningful unit in a word. An Amharic root is a sequence of base characters. A collection of phonemes or sounds creates a word, which can be as simple as a single morpheme or contain several of them.

Another problem of the language is that there are different ways of writing a single word due to different reasons. One reason for this can be regional dialects that can impact word formation in the basic level where the words are more likely to be written following their spoken form; "ሃጂ" vs. "ሂጅ", "ዓጠ" vs. "ዓፄ" , etc. Another one is, in Amharic there are many ways of writing loan words, i.e. words that are taken from foreign languages. For example, the word "Computer" can be written as ኮምፒዩተር, ኮምፒውተር.

4.2 System Design

The construction model of our semantic network whose architecture is shown on Figure 4.1 involves:-

- ✓ Text analysis and creation of index file for the collected news text corpus
- ✓ Formation of term vectors using WordSpace model based on the index file
- ✓ Searching the WordSpace, which is collection of term vectors, to extract concepts (semantically related ones) for a given Amharic WordNet synset
- ✓ Finally extract relations that exist between those concepts using intervening words pattern, which are extracted from the corpus using pairs of concepts from Amharic WordNet of specific relation 'R', again from the corpus.

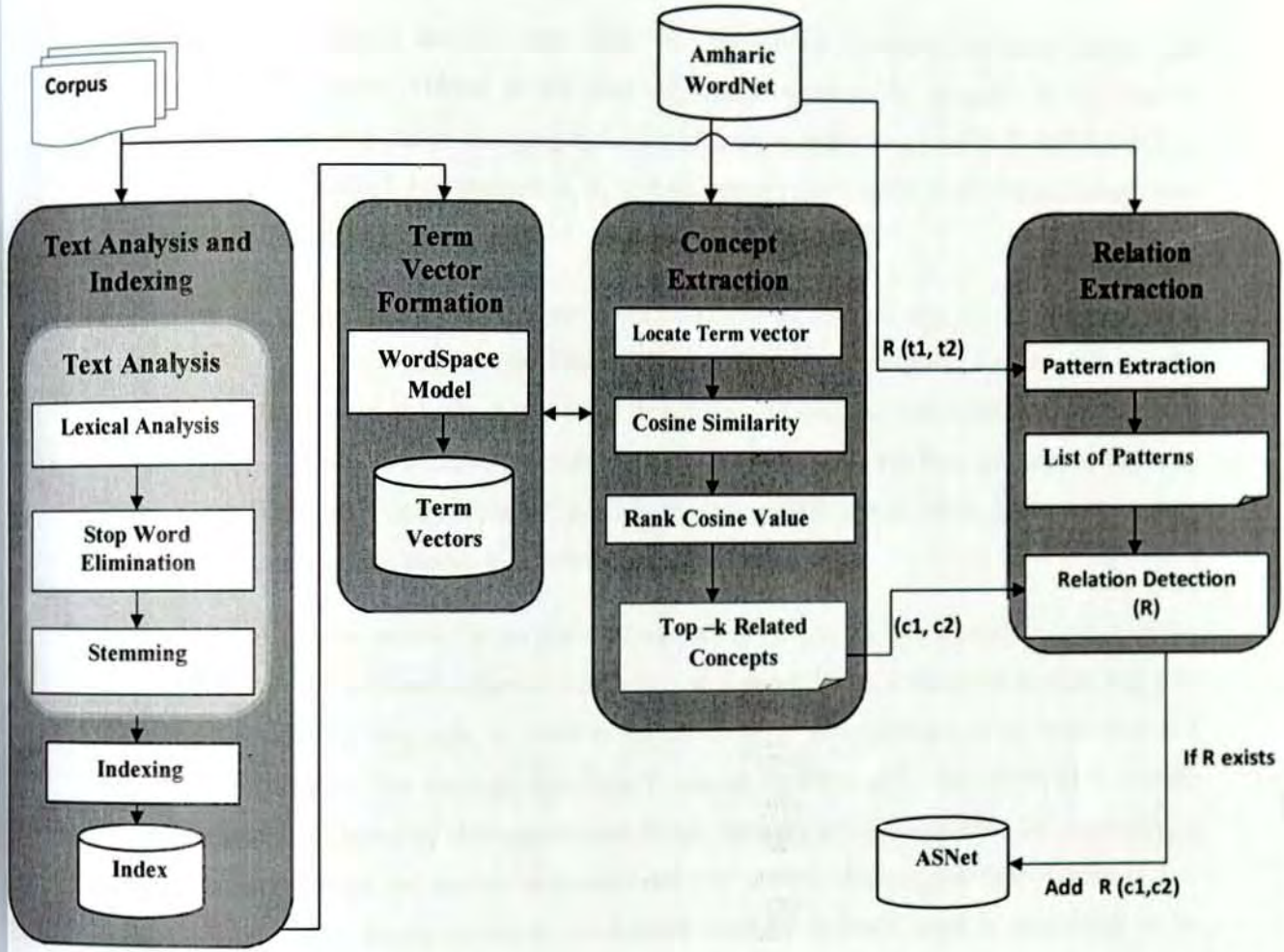


Figure 4.1. Architecture of ASNet

4.2.1 Text Analysis and Indexing

The process of corpus analysis starts with the removal of non-letter character tokens and normalization of them. Tokens in this case are words separated by space(s). In our system, normalization is a process of changing alphabets of same sound (phone) but different symbols in to one common symbol. For example ሀ, ሐ, ኀ, ሃ are normalized to ሀ, for every word having those alphabets in the corpus.

The next activity in the text analysis part is the process of removal of stop words whose meaning is less relevant to the document specifically and for the corpus in general. No one agrees on the idea that a given word is really a stop word or not. But we take the commonly known words in which average numbers of researchers decide them as stop words. We have got around 130 stop word lists from [18]. For example ነው, አቶ, እና are most known stop words in Amharic language. The list of stop words is shown in appendix I.

Stemming is the next activity in the text analysis part of the system. It is a process of finding the stem which is the last unchangeable morpheme of a word. Since Amharic is morphologically rich and inflectional language, we need to have a common representation of all inflection of a word in the corpus. The stemmer algorithm developed by Alemayehu and Willet [22] is used. The stemmer developed by Alemayehu and Willet removes affixes iteratively by employing a minimum stem length and context sensitive rules with prefixes are removed before suffixes. For the minimum stem length condition, the shortest form of Amharic word is considered to be composed of two syllables. Therefore, the stemming is applied only to those words whose length is more than two syllables. The algorithm removes the suffix and the last character of the remaining word is changed to *Sads* (the sixth order of Amharic character or syllable).

The algorithm has some problem of stemming verbs. The algorithm cannot properly stem different inflections of a verb as a result we may get more than one morpheme as a stem of inflections of a single verb. This means, different inflections of a word may have different stems, but actually should be one stem for those inflections. After removing the suffixes, the algorithm changes the last character to *sads* for all the words need to be stemmed in the corpus. For surface words with suffix beginning by a vowel, the algorithm changes the last character to *sads* correctly but this process will not apply to surface words with suffix beginning by a consonant. For example, the words ኢትዮጵያ-ውያን with suffix "ውያን" and ባንክ-ኦች with suffix "ኦች" are

stemmed to ኢትዮጵይ and ባንh respectively. Since the suffix "ኦች" starts with 'ኦ' which is vowel, the word ባንhኦች is stemmed correctly. "ውያን" is a suffix that starts with 'ው' which is a consonant and the word ኢትዮጵያውያን did not stem correctly. Since we are interested in noun concepts, we have modified the implementation of the algorithm to treat the way the last character changes to *sads* correctly for those words having suffixes starting with consonants.

Once the documents in the corpus are analyzed, the last process is indexing it to increase efficiency of searching term frequency and fit the data for WordSpace generation. We have used the TF-IDF term weighting algorithm to store only highly weighted (relevant) terms to the document in specific and to the corpus in general. So, the size of the index decreases as the low weight terms are removed and will not be stored in the index file. The term is added to the index file along with its statistics. The corresponding statistics about the term is document frequencies and position of the term within the documents.

The algorithm to index the corpus is adopted from apache lucene indexer [15]. Apache lucene is a full fledged search engine and is available freely on the web. The indexer of this search engine uses TF-IDF term weighting algorithm for filtering relevant terms. The inputs to the indexer are the main directory AMCORPUS having all sub directories SUBDIR or documents and the folder INDEX to hold the indexed files. If a FILE encountered is a document, then the algorithm applies text operations or analysis (i.e., lexical analysis, stop word elimination, and stemming) and adds the document to the index file. Listing 4.1 shows the pseudo code of the algorithm to index the corpus and store as index files.

```
Inputs:
AMCORPUS=corpus directory
INDEX= indexed file directory
    1. Open AMCORPUS
    2. For each SUBDIR in AMCORPUS
    3.     Open SUBDIR
           For each FILE in the SUBDIR:
                If FILE is Directory
                GOTO 3
                Else
                Apply text operation on FILE
                Add FILE to INDEX
```

Listing 4.1. Algorithm for Creating Index File

4.2.2 Term Vector Formation

As we said in Chapter Two, a term vector is a sequence of term-weight pairs. But the weight of the term in our case is the co-occurrence frequency of a term with the other term in a document. In the vector space model of information retrieval, documents are modeled as vectors in a high-dimensional space of many terms. The terms are derived from words or phrases in the document and are weighted by their importance within the document in particular and within the corpus in general.

The process of creating term vectors uses the index file. Once we have the index file, it is possible to get and use the list of all the terms in the entire index. Then for the terms which we want to consider for our similarity computation, we extract the co-occurred terms to form their term vectors in the WordSpace model. The vector gives us an array of terms along with their co-occurrence frequency for each term in the index file.

From the index file, it is possible to map the index to term-context (term-document) matrix where the values of the cells of the matrix are the weighted frequency of terms in the context (document). The WordSpace model is used to create term vectors semantically from this matrix by reducing the dimension of the matrix by using random projection algorithm [34]. Since a matrix has many cells and most of the cells are full of spares data, there must be a means to reduce the size and eliminate the spares cells of the matrix. Singular value decomposition (SVD) is a type of a random projection algorithm and is used to reduce the dimension and eliminate the sparse data of the matrix [34]. The need to change the term vector to matrix is that the input to the WordSpace model is a matrix. At the end, the WordSpace contains the list of term vectors found from the corpus along with co-occurrence frequencies of each term. The pseudo code of the algorithm to develop the WordSpace model is shown in listing 4.2 [26]. The inputs for the algorithm are the index file which is indexed before.

The algorithm maps the index file to a term-document matrix so that we can create document and term vectors easily from the matrix. For each document in the column of the matrix, the algorithm creates random document vector to reduce the dimension of the matrix by using random projection algorithm [34]. Once the random document vector for each document in the matrix is created, we can calculate the frequency of each term in the document and multiply by the random document vector value. This gives us a value for term vector in that document.

At the final stage, the term vectors are computed and normalized based on the term frequency within each document and the random vector produced as an output from the Random Projection algorithm. The WordSpace is composed of the normalized term vectors in the document collection.

Term vector consists of numbers which indicate the co-occurrence frequency of the term with the remaining terms. For example, if the co-occurrence frequencies of the term ኢትዮጵያ (Ethiopia) with አማራ (Amhara), አሮማያ (Oromia), ኢፌዴሪ (Ethiopia Federal Democratic Republic) and መንግስት (government) are 20,10,15 and 3 respectively, then the term vector TV={20,10,15,3}. So, after the creation of the term vector, some mathematical calculations like Cosine function can be performed on it.

```
Input:
INDEX=folder containing index files
DIM=dimension of each term vector
Output:
TERMVECTORS=set of term vectors with dimension Dim
Initialize:
TERMVECTORS={}
DIM=200 //default dimension of each vector
   1. Open INDEX
   2. Create term-document matrix, MATRIX from INDEX
   3. For each COLUMN in MATRIX: //for each document
            Create basic random vector, RAN_DOC_VEC
            For each TERM in CULUMN AND dimension<=DIM: //for each term on the row
                  FREQ = frequency of the term within the document
                  TERM_VEC={} //term vector
                     TERM_VEC = TERM_VEC + (FREQ * RAN_DOC_VEC)
                  Normalize TERM_VEC
                  Add TERM_VEC to TERMVECTORS
Return TERMVECTORS
```

Listing 4.2. Algorithm for Creating WordSpace

4.2.3 Amharic WordNet

To construct ASNet automatically from free text corpora using WordSpace model, we still need some information as background knowledge for the system so that other unknown relation instances can be extracted. Amharic WordNet is constructed as a small knowledge base. The basic relation used to construct Amharic WordNet is "synonymy" which is used to group semantically related terms together. Amharic WordNet is composed of 890 single word terms (all are nouns) grouped into 296 synsets (synonym groups) and these synsets are representation of concept of the terms in the group. We chose noun concepts because most relation types are detected between nouns. Verbs and adverbs are relation indicators which are used to show relations between nouns. Synsets are further related with each other by other three relations called type-of (hierarchy), part-of (part-whole) and antonym. Figure 4.2 shows relations that are incorporated in Amharic WordNet. The basic need of Amharic WordNet in this thesis is to set different seeds for a specific relation. Once we prepare sets of seeds from this WordNet as input to the ASNet, we can extract the patterns which indicate how these pairs of seeds exist in corpus. The way these pairs of concepts exist in corpus can tell us more about other concept pairs in corpus.

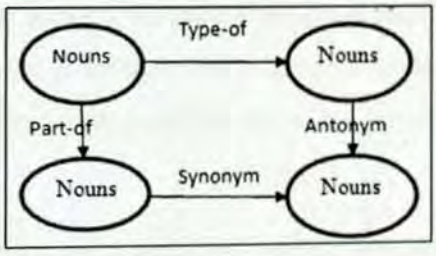


Figure 4.2. Relations in Amharic WordNet

For example, the way the pair of terms {ኢትዮጵያ(Ethiopia), አፍሪካ(Africa)}exists in the corpus can tell us that the pair of terms {ኬንያ(Kenya), አፍሪካ} can exist in same way as the former pairs. The patterns we extracted between pair of terms {ኢትዮጵያ, አፍሪካ} can be used to extract the relation between other countries like ኬንያ with that of አፍሪካ. Table 4.1 shows example of Amharic words organized as synsets (concepts) in Amharic WordNet. Each concept has a unique identification number. The number is used in Amharic WordNet to pair concepts with different relations.

Table 4.1. Synsets in Amharic WordNet

ID	Words in Synset
0001	ኢትዮጵያ, አቢሲኒያ, ኢፊድሪ
0002	ሀብት, ንብረት, ገንዘብ
0003	ኢኮኖሚ
0004	አገር
0005	አለም, ፕላኔት
0006	አፍሪካ, አፍሪካ ሀብረት
0007	ፕሮጀክት, እቅድ, ፖሊሲ

The above concepts are further connected by meronym or "part of" and hyponym or "type of" relations in Amharic WordNet. Table 4.2 shows the two relations in between the above concepts. The pair of entries in the table tells us that the first concept is "type of" and "part of" the second concept respectively.

Table 4.2. Sample Relations in Amharic WordNet

Type of Relation	Part of Relation
{0001,0004}	{0001,0005}, {0003,0001}, {0002,0005}, {0003,0006}
{0002,0003}	{0001,0006}, {0007,0001}, {0002,0006}, {0004,0005}
	{0007,0004}, {0002,0001}, {0003,0004}, {0004,0006}
	{0002,0003}, {0002,0004}, {0003,0005}, {0006,0005}

4.2.4 Concept Extraction

Concept extraction is the main and essential process in this thesis. As we said in chapter one, concepts can be found at document, paragraph, sentence, phrase or word level. For example, the phrase ትምህርት ሚኒስተር (Ministry of Education) is a combination of two words. But this multi word term stands for a single concept. Splitting this phrase leads to get another two concepts, ትምህርት (Education) and ሚኒስተር (Minister), that is found at word level. Even though Amharic concepts can exist on different levels described above, we are interested to extract it at word level.

WordSpace model is used to extract semantically related concepts for a seed term of Amharic WordNet. WordSpace model is used to create collection of term vectors in which each term vector contains different related words along with their co-occurrence frequencies. For a concept from Amharic WordNet as input to WordSpace, related concepts are extracted by computing the cosine value between the term vector containing this concept and the remaining term vectors of the WordSpace model.

For example, consider the synset (concept) {ኢትዮጵያ, አቢሲኒያ, ኢፌዴሪ} to mean {Ethiopia} from Amharic WordNet, we need to extract other related concepts of this synset from the model. First, locate in which term vector this concept is found the WordSpace. Assume the co-occurred terms, {ኢትዮጵያ, አፍሪካ, ኔፓድ, አለም, ንግድ, ከልል, ...} of {አገር} maped to a term vector "TV" of value {5, 7, 3, 10, 20, 8, ...} respectively, the concept from Amharic WordNet used as input is now located in this term vector "TV". This is because the word {ኢትዮጵያ} which is a member of the sysnet is found in the term vector "TV". The numbers show that how many times the word {አገር} co-occurred with the listed words. The next activity is to calculate the cosine value of the

angle between the term vector "TV" and the remaining term vectors in the WordSpace. The cosine value is always between 0 and 1. If the result is nearly 1, then the words in each term vectors (cosine calculated between) are semantically related. If the cosine value is nearly 0, then there is no relation between those words of the term vectors.

Since the collection of the term vectors in the WordSpace is many in number, it is a must to limit how many related terms need to be extracted to a certain input synset. Ranking the related terms to the given sysnet based on the cosine value is essential to limit the dimension of the result.

When related concepts are extracted, we have considered the following points:

- We assumed that most of the concepts in the corpus are nouns. We have extracted from the model only nouns by using the manually tagged data as reference. Different word categories can not be related at any way.This increases the accuracy of extracting related concepts.
- The corpus under consideration is composed of news documents which used the system to train in a domain independent fashion. So, we had incorporated around 4325 news items from Ethiopian news agency to get possible frequency of terms.
- We have developed sample Amharic WordNet noun entries to be a base for the system to aquire new knowledge.
- The way we used the WordSpace model is to reduce the search space so that instead of searching a relation betwean every two terms of the WordSapce, we can detect a relation betwean every two terms from those related concepts which are results of searching the model using Amharic WordNet seed terms.

In general, in section 4.2.2, we have shown how the WordSpace which contains term vectors is developed for the corpus we used. In the algorithm shown in Listing 4.3, for each synset (concept) in the Amharic WordNet, we search the WordSpace to get semantically related concepts using cosine similarity measuring algorithm. We first locate in which vector the given synset member is found in the WordSpace. Once the term vector is located, we calculate the cosine similarity of the vector which contains the synset with all other vectors in the WordSpace.

After cosine similarity is computed, we rank the cosine values in decreasing order for selection of top-k number of related concepts for the given synset where 'k' is our threshold to extract how many number of related concepts to be extracted.

Inputs:

SYNSET=AmW terms in each synset

WORDSPACE=Wordspace model to search term vectors in.

K= a treshold that how many term vectors should be selected

Output:

CONCEPTS= set of semantically related concepts with its cosine value

Initialize:

CONCEPTS={}

1. Accept SYNSET having synonym terms

2. Normalize and stem the terms in the SYNSET

3. Open WORDSPACE

4. Locate a term vector, TERM_VEC that correspond to the SYNSET

5. For each term vector, $TERM\_VEC_i$ in the WORDSPACE:

$$COS\_SIM = \frac{\sum_{i=1}^{n} TERM\_VEC * TERM\_VEC\ i}{\sqrt{\sum_{i=1}^{n} TERM\_VECx^2 * TERM\_VECx^2\ i}}$$

Where n is number of term vectors in the model and COS_SIM is cosine value calculated

6. Rank cosine values

6. select top k term vectors based on their COS_SIM and add to CONCEPTS

7. Return CONCEPTS

Listing 4.3. Algorithm for Generation of Semantically Related Concepts

4.2.5 Relation Extraction

Relation extraction is the process of clustering entities into some given relationship categories within textual documents. We can say that relation extraction is a process of extracting entities and relations among them from text documents. In this thesis, an entity stands for concepts. Examples of concepts are Ethiopia, Africa and country. Examples of relations are "part of" a region and "types of" country. For example, Ethiopia is a country and is part of Africa.

Detecting semantic relations in text is very useful in both information retrieval and question answering because it enables knowledge bases to be powered to score passages and retrieve candidate answers. As we said before in chapter three, to extract semantic relations from text, three types of approaches have been applied. The first is rule based methods (knowledge based) which employ a number of linguistic rules to capture relation patterns. The second is feature based methods (Corpus based) which transform relation instances into a large amount of linguistic features like semantic features, and capture the similarity between these feature vectors. The last is combining different approaches (hybrid approach) so that extraction performance can be improved.

Hybrid approach uses a very small number of seed instances or patterns to do bootstrap learning. These seeds are used with a large corpus to extract a new set of patterns, which are used to extract more instances, which are used to extract more patterns, in an iterative fashion.

ASNet is designed based on the third approach which needs minimum human supervision and acquires more information from free text corpus. This is the so called semi supervised approach. For example, given a small seed set of {አገር (Country), አህጉር (Continent)} pairs, the approach we applied perform the following minimum steps:

1. Use the pairs to extract some data in a corpus
2. Induce patterns from the extracted data.
3. Apply the patterns to the corpus in general to get new set of {አገር (Country), አህጉር (Continent)} pairs, and add to the seed set.
4. Return to step 1, and iterate until convergence criteria is reached

In general, using Amharic WordNet entries, intervening words patterns for a specific relation are extracted from the corpus. For each pair of concepts (C1, C2) of which we know the relationship

contained in Amharic WordNet, we send the query "C1" +"C2" to the corpus. The returned text snapshot is processed to extract all n-grams ($2 \leq n \leq 7$) that match the pattern "C1 X * C2", where X can be any combination of up to five space-separated word or punctuation tokens. Thus "C1 X * C2" where "C1"and "C2" are replaced with variables is a pattern extracted from the corpus using concept pair (C1, C2) from Amharic WordNet of specific relation "R".

For instance, assume the Amharic WordNet contains the concepts "ኢትዮጵያ" and "አማራ" with ኢተዮጵያ being a hypernym of አማራ. The method would query the corpus with the string, "ኢተዮጵያ" +"አማራ". Let us assume that one of the returned text snapshot is "...በኢትዮጵያ ከሚገኙ ከልሎች መካከል አማራ አንዱ ሲሆን...". In this case, the method would extract the pattern "በኢትዮጵያ ከሚገኙ ከልሎች መካከል አማራ" .This pattern would be added to the list of potential hypernymy patterns list with "ኢተዮጵያ" and "አማራ" substituted with matching placeholders, like "var1 ከሚገኙ ከልሎች መካከል var2". The algorithm to extract intervening words patterns is shown in Listing 4.4.

```
ExtractPatterns(D, C)
Inputs:
C=set of pair of concepts from AmW for specific relation decided first
D=corpus
Output:
P= set of patterns for the given relation R
  Initialize: P={}
    1. For each pair (A,B) in C
       2.For each file F in D
            3.For each sentence S in F
                 CONTEXT= Entervn(S,A,B)
                 P=P+ CONTEXT

    Return P
```

Listing 4.4. Algorithm for Extraction of Patterns

Of course, the set of patterns extracted in this way is too large to be used directly for relation extraction. Therefore, we rank the patterns according to their ability to distinguish between the types of relationships we are interested in. To do this, we need to get the frequency of the patterns from the corpus and ranked based on their frequency. So, highly frequent pattern is added to the list of patterns for that specific relation.

To extract a set of patterns using concept pairs from Amharic WordNet for the specified relation R, we should get the intervening words which are n-gram terms between the given pair of concepts in the sentence. To do this we have an algorithm shown in listing 4.5. Patterns extracted in the above way are used for extraction of new concepts which are not found in the Amharic WordNet. Concepts which have the same pattern in the corpus with seed concepts of Amharic WordNet can be extracted as new one. For instance, using the pattern "var1 ከሚገኙ ከልሎች መካከል var2" and searching on the corpus, we may get text snapshots like "በኢትዮጵያ ከሚገኙ ከልሎች መካከል ኦሮሚያ" and "በኢትዮጵያ ከሚገኙ ከልሎች መካከል አፋር". From these text snapshots, ኦሮሚያ and አፋር are new concepts extracted from the corpus and are hyponym of the term ኢትዮጵያ.

```
Entervn(S, A, B)
Inputs:
S=sentence which is considered as array of words
A=first Concept in the pair
B=second Concept in the pair
Output:
PHRASE= n-gram terms (2<n<=7) betwean A and B in the sentence S
Initialize:
PHRASE={}
    1. For i=1 to n-2
            PHRASE=PHRASE+S[i]
       Return PHRASE
```

Listing 4.5. Algorithm for Extraction of Intervening Words

Once the patterns are extracted, the final step is to detect if there is a relation R betwean every pair of concepts extracted from the WordSpace. The relations that are detected automatically are

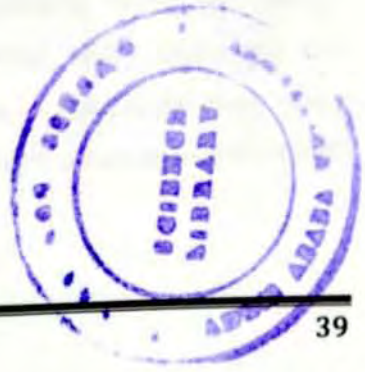
meronym (part-whole) and hierarchy (type-of). These are potential relations between noun concepts in any corpus. Consider the phrase "ከጥራጥሬ አይነቶች መካከል አንዱ የሆነው አተር... "extracted from the corpus using concept pair (አተር,ጥራጥሬ) of Amharic WordNet, where አተር is type of ጥራጥሬ. This pattern is added to the list of hyponym (type-of) patterns with replacing the terms ጥራጥሬ and አተር by placeholders or variables. So the pattern becomen "var1 አይነቶች መካከል አንዱ የሆነው var2".

Now, we need to detect a type-of relation between pairs of concepts extracted from the WordSpace. Assume those concept pairs are for example (ምስር,ጥራጥሬ) and (በቆሎ,ጥራጥሬ). To detect the type-of relation between these concept pairs using the above pattern, first we replaced the varible var1 with ጥራጥሬ and var2 with ምስር, then we searched the corpus if the phrase "ከጥራጥሬ አይነቶች መካከል አንዱ የሆነው ምስር" exists. If it exists, then we get a new concept pair (ምስር,ጥራጥሬ) with type-of relation and will be added to the network. The same is true for the pair (በቆሎ,ጥራጥሬ) in which ጥራጥሬ will replace var1 and በቆሎ will replace var2. The corpus will be searched with the phrase „ከጥራጥሬ አይነቶች መካከል አንዱ የሆነው በቆሎ". If it occurs, then the concept pair (በቆሎ,ጥራጥሬ) will be added to the network in which each concept is a node and the link is the relation between the concepts. The algorithm shown on Listing 4.6 is used to extract a relation between pairs of concepts, which came from the WordSpace.

The inputs of the algorithm are set of pairs {R (C1,C2)} and the corpus, where R is a specific realtion and C1, C2 are concept pairs. The concept pairs are extracted from the WordSpace using patterns of text in corpus. The corpus is composed of domain independent free text data. It is a collection of news text documents gathered from Walta Information Center and Ethiopian National News Agency. The output of the algorithm is a prolog style text representation as that of the input concept sets.

```
ExtractRelations(D, Patterns, C)
Inputs:
C=set of pair of concepts from the RELATED CONCEPTS group
D=corpus
Patterns=list of patterns extracted in the pattern extraction proccess
PHRASE={}, to store list of modified patterns
Output:
SemNet=list containing pairs of concepts with relation R as a form of R(C1, C2)
1. For each pair (A, B) in C
    1.1 For each p in Patterns
        Pat=A+p+B
        PHRASE= PHRASE+pat
2. For each phrase in PHRASE
    Count=0
    A=phrase[0]
    B=phrase[size(phrase)-1]
    2.1 For each file F in D
        2.2.For each sentence S in F
            If phrase exists in S
              Count++
          If count>=3 //treshold of pattern to detect relation R
            Add pair R(A, B) to SemNet
            break
Return SemNet
```

Listing 4.6. Algorithm for Relation Extraction



4.2.6 ASNet

ASNet consists of a set of concepts and a set of useful and important relationships called "Synonym", "Part of" and "Type of". Semantic networks not only represent information but also facilitate the retrieval of relevant facts. For instance, all the facts about the concept {ኢትዮጵያ (Ethiopia)} are stored with pointers directly to the node representing {ኢትዮጵያ (Ethiopia)}. Another example concerns the inheritance of properties. Given a fact such as "አ ገር ሁሉ መንግስት አለው (a country has a government)", ASNet would automatically conclude that "ኢትዮጵያ መንግስት አላት (Ethiopia has a government)" given that ኢትዮጵያ አገር ነች (Ethiopia is a country). Figure 4.3 depicts this scenario.

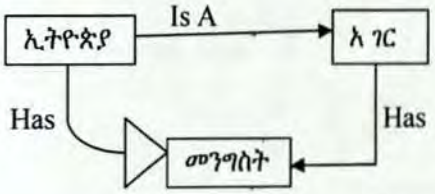


Figure 4.3. Inheritance of Properties in ASNet

ASNet acquires new concepts over time and connects each new concept to a subset of the concepts within an existing neighborhood, whenever new text document is processed by the system. Our growing network model is not intended as a complete model of semantic development, but to show how specific relations can be extracted and connected between concepts of the given corpus.

ASNet is developed to hold entries as a form of first order predicate calculus in which the predicate is the relation and the arguments are concepts. All the concepts are noun entries and the relations are synonym, type-of and part-of. Consider the entry, Hypo (አዲስ አበባ,ከተማ), to ASNet; The first argument, አዲስ አበባ(Addis Ababa) is hyponym of the second argument, ከተማ (town). The same is true for the entry, part-of (ኢትዮጵያ, አፍሪካ) in which the first argument, ኢትዮጵያ(Ethiopia) is part of the second argument, አፍሪካ (Africa). The predicates, "Hypo" and "part-of" are relations between concept pairs, (አዲስ አበባ,ከተማ), and (ኢትዮጵያ, አፍሪካ) respectively. For example, the entry Hypo (አማራ,ክልል) to ASNet is shown as አማራ ⟶ ክልል in the graph where ⟶ shows type-of relation. The same is true for entry Part (ኢትዮጵያ,አለም) shown as ኢትዮጵያ ⟶ አለም where ⟶ shows part-of relation.

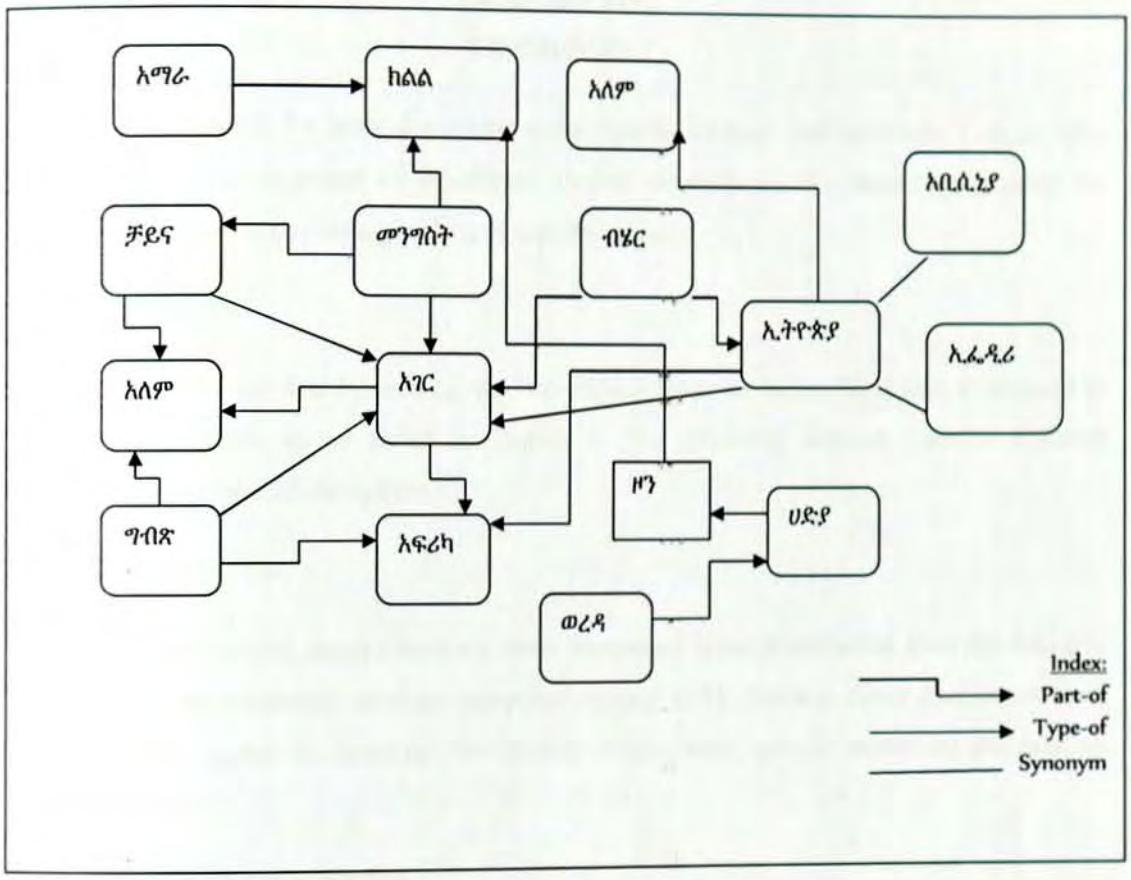


Figure 4.4. Sample Semantic Network in ASNet

CHAPTER FIVE
EXPERIMENT

ASNet is implemented for news documents using Apache Lucene[1] and Semantic Vectors APIs for indexing and development of WordSpace model, respectively. We have implemented the pattern and relation extraction algorithms to test the model.

5.1 Implementation

ASNet is implemented first by creating the WordSpace from the index file which is mapped to term-document matrix as we stated in Chapter 4. The following sections describe different implementation phases of the system.

5.1.1 Corpus Collection

For corpus establishment, tagged Amharic news documents were downloaded from the Amharic NLP[2] website for extraction of noun concepts. Around 4325 Amharic news documents from Ethiopian news agency are used for WordSpace development, pattern extraction and relation detection processes.

5.1.2 Text Analysis and Indexing

All the text operations are performed by utilizing the "AmharicAnalyzer" class which is inherited from Lucene's Analyzer class. Basically, the "AmharicAnalyzer" class discards non-letter characters, normalizes Amharic characters for spelling inconsistency, eliminates stop words, and stems terms. The source code of the Amharic stemmer written by Tessema (2007) is used to stem terms [22].

Finally, Apache lucene indexing program based on TF-IDF term weighting is run to index all of the documents in the corpus [15]. The index file is saved in binary file format on the disk for efficiency reasons. The index file is composed of more than 28,000 index terms (key words) extracted from around 4325 documents.

[1]Available at: http://lucene.apache.org/core/ (Date Accessed 12/04/2012)

[2]Available at: http://nlp.amharic.org/resources (Date Accessed 10/02/2012)

5.1.3 WordSpace Model

Semantic Vectors[3] API package version 3.2 is used to create the WordSpace which is composed of term vectors [10]. Using random projection algorithm, the vector dimension is reduced to 200 which is the default value the API used.

5.1.4. Concept Extraction

In order to generate related concepts for a given seed term of Amharic WordNet, the WordSpace is opened and searched with that term. This is for the purpose of space reduction in searching of semantic relations.

Table 5.1. Example of Terms and Their Semantically Related Concepts Extracted

Term	Related Concepts	Non Related Concepts
ኢትዮጵያ	መንግስት፣ ንግድ፣ አገር፣ ህዝብ፣ ድሬዳዋ፣ ቢተክርስትያን፣ ሚኒልክ	ውስጥ፣ መስማማት፣ የተለያዩ
አፈሪካ	አለም፣ አገር፣ መንግስታት	ደቡብ፣ አባል፣ መር
ብሄረሰብ	ብሄር፣ ማንነት፣ አኝዋክ፣ ቁንቁ	ሀምዛ፣ መከበር
መንግሰት	ኢትዮጵያ፣ ህዝብ፣ ህግ፣ ፖለቲካ	መለስ፣ ቤት፣ ድጋፍ፣ ችግር፣ አስተያየት

5.1.5 Relation Extraction

Concepts related by meronym (part-whole) and hierarchy (type-of) relations of Amharic WordNet are used for extracting patterns from the collected news documents. The pattern extraction algorithm is implemented to extract intervening words patterns between given pairs of concepts of Amharic WordNet for meronym and type-of relations. Highly frequent patterns are added to the potential meronym and type-of patterns list. The relation detection process utilizes the extracted patterns to detect if there is a meronym and/or type-of relations between two pairs of concepts, extracted from the WordSpace.

[3]Available at: http://semanticvectors.googlecode.com (Date Accessed 12/04/2012)

5.3 Evaluation of ASNet

Evaluating the network that automatically developed is a very difficult task. The nodes (concepts) that are related with many relations in the corpus are also many in numbers and cannot be evaluated easily whether the relation between each and every node is correct or not. The main idea behind creating the semantic network automatically is to extract and build huge knowledge base with minimum time usage trading some precision loses in the construction process. Here, we need to show the different intermediate results along with its evaluation done manually.

Firstly, we have tested the system with seven terms selected manually as initial seed terms to extract semantically related concepts from the WordSpace, developed from 1065 tagged news corpus, using cosine algorithm. For each seed term, we have retrieved the top 20 related concepts from the WordSpace. Among the 20 related concepts, we selected noun concepts using the same POS tagged corpus that is previously used for the development of the WordSpace. The result is shown in table 5.2.

Table 5.2. Semanticaly Related Terms for the Given Seed Terms

Term	Semanticaly Related Terms	Not Semanticaly Related Terms
አፍሪካ	አፍሪካ፣አለም፣አገር፣መንግሰት፣ካርተር	ሀብረት፣አቀፍ፣ደቡብ
ኢትዮጵያ	ኢትዮጵያ፣አገር፣መንግሰት፣ንግድ፣ድሬዳዋ፣ህዝብ፣ቤተክርስትያን	ውሰጥ፣መሰማማት
ኔፓድ	ኔፓድ፣አለም፣ማህበረሰብ፣ካናዳ፣አጋር፣ጂኤይት፣ኢንቨስትመንት፣ፓርትነርስፕ	ጆንቨረ፣ማኔጂንግ
መንግስት	ኢትዮጵያ፣መንግሰት፣ህዝብ፣ህግ፣ፖለቲካ፣መለስ፣ቤት፣ችግር	ድጋፈ፣አስተያየት
ህዝብ	ህዝብ፣መንግሰት፣ቤት፣ችግር፣አገር፣ልማት	መዋቅር፣ፎርም
ብሄረሰብ	ብሄረሰብ፣ብሄር፣ማንነት፣ወራብ፣አኙዋክቃንቃ	ሀምዛ፣ኑዌር፣ወራብ
አማራ	አማራ፣ወሎ፣ክልል፣አሮሚያ	ምግብ፣ቅነጅት

In the second experiment, we have increased the number of documents in the corpus to have possible frequency of terms. Again we used the same tagged dataset as reference to extract only noun concepts in the corpus. In this case additional 2421 documents were added to the corpus for possible frequency and co-occurrence analysis of terms. We have used around 20 pairs of seeds of Amharic WordNet to extract patterns for part-of (meronym) relation. The seeds are also linked by part-of relation in Amharic WordNet. The pairs of seed terms are shown in table 5.3.

Table 5.3. Part-of Relation Seeds for Pattern Extraction

ዞን:ክልል	ህግ:መንግስት	ወረዳ:ዞን	ሃብት:አገር
አገር:አለም	ቀበሌ:ወረዳ	ህዝብ:አገር	አገር:አለም
ከተማ:አገር	ክልል:ኢትዮጵያ	ተማር:ዩኒቨርሲቲ	ኒዮርክ:አሜሪካ
ትግራይ:ኢትዮጵያ	ኢኮኖሚ:ኢትዮጵያ	መንግስት:አገር	መንግስት:አማራ
ኢትዮጵያ:ኔፓድ	ኢትዮጵያ:አፍሪካ	መንግስት:ኢትዮጵያ	ብሄረሰብ:ኢትዮጵያ

Terms in the pair shown in table 5.3 are separated by colon (:). This is to mean the concept of the first term is part of the concept of the second term. These pairs of seed terms are manually selected after the corpus is indexed and the selection criteria were just based on their frequency in the corpus.

We got more than 103 text snapshots as intervening words pattern using these pairs of seed terms. Using the patterns, we further detect around 35 correct part-of relations from among 57 newly extracted pairs. Correctly extracted pairs of concepts for part-of relation are shown in Table 5.4. The concept of the first term is part of the concept of the second term in the pair. Since the corpus we have used in the experiment was compiled on news documents, the correctness of the pairs is manually traced irrespective of the domain of the corpus. So, the correctness is relatively conceptuall based on the meaning of the available news items.

Table 5.4. Extracted Concept Pairs with Part-of Relation

(ስራ,ህዝብ)	(ቤት,አገር)	(ህግ,ፖለቲክ)	(ስር,አገር)
(አገር,ህዝብ)	(ኢትዮጵይ,አፍሪክ)	(አገር,አለም)	
(ችግር,አገር)	(ህዝብ,አገር)	(ህዝብ,ኢትዮጵይ)	
(ህዝብ,ፖለቲክ)	(አገር,አፍሪክ)	(ኢትዮጵይ,አለም)	
(መንግስት,አገር)	(ህዝብ,መንግስት)	(ስር,መንግስት)	
(ህግ,መንግስት)	(መንግስት,ህዝብ)	(ችግር,ህዝብ)	
(ኢትዮጵይ,መንግስት)	(መንግስት,ኢትዮጵይ)	(ህዝብ,ልማት)	

To show the network of the above concept pairs, we have sketched a graph containing concepts as nodes and relations as edges. The output of the above experiment is shown as a graph in Figure 5.1.

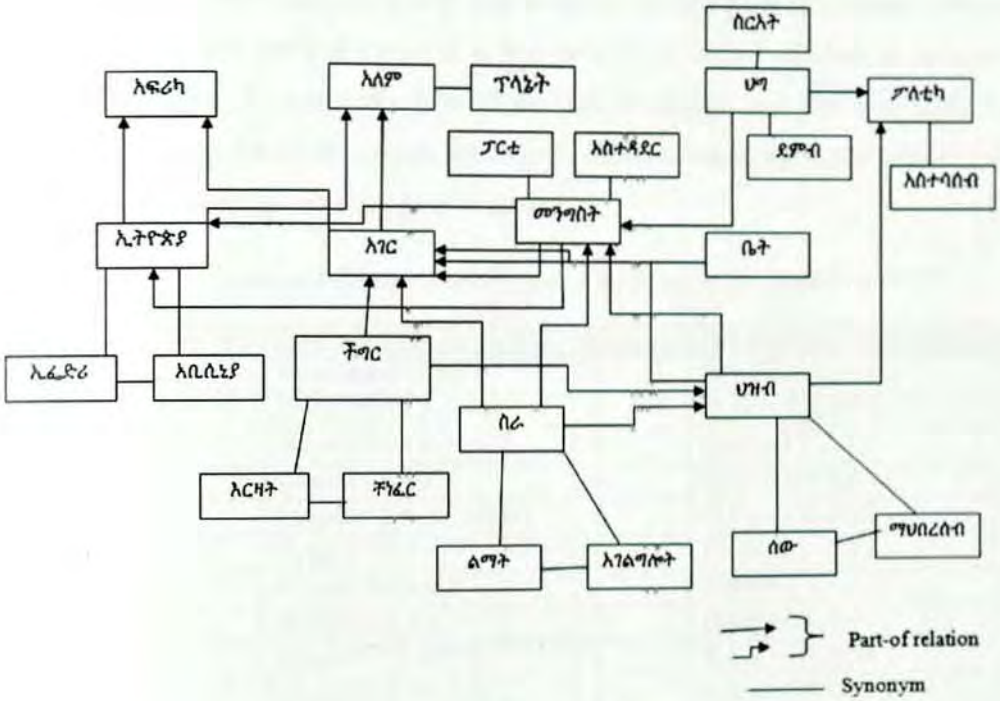


Figure 5.1. Extracted Concept Pairs with Part-of Relation Shown as Graph

The last experiment was done using 4325 news documents along with the POS tagged news documents as reference for noun concept extraction purpose. In this experiment, we have used 21 seed terms which are highly frequent in the corpus. Using these terms, we have extracted 588 pairs of noun concepts from WordSpace using cosine algorithm. The 21 seed terms are shown in Table 5.5. In this experimentation phase, we consider both part-of and type-of relations.

Table 5.5. Seed Terms from Amharic WordNet

አ ፍ ሪ ካ	ኢ ት ዮ ጵ ያ	መ ን ግ ሥ ት	ብ ሄ ረ ሰ ብ	አ ማ ራ	አ ገ ር
ኢ ኮ ኖ ሚ	አ ለ ም	ድ ር ጅ ት	ከ ል ል	ህ ግ	
ል ማ ት	ፖ ለ ቲ ካ	ት ም ህ ር ት	ወ ረ ዳ	ፕ ሬ ዚ ዳ ን ት	
ከ ለ ብ	ስ ፖ ር ት	ሱ ዳ ን	አ ት ሌ ት	ተ መ ድ	

To detect the part-of and type-of relations between each pair of concepts we extracted from the WordSpace, we have prepared 137 and 151 seed pairs for type-of and part-of relations respectively. List of pairs of seed terms for the two relations are shown in appendix II. Using these seed pairs as input to the system, we have extracted intervening words patterns between each pair from the corpus.

The patterns that we feed manually in turn used to extract 58 new pairs of concepts having type-of relations and 76 new pairs of concepts having part-of (meronym) relations in between. The newly extracted pairs of concepts for type-of and part-of relations are shown in Table 5.6 and Table 5.7 respectively. Under the remark column, '√' shows correctly extracted pairs of concepts and '×' shows not correctly extracted pairs of concepts.

Table 5.6. Extracted Concepts with Type-of Relation Represented by FOPC

Concept pairs	Remark	Concept pairs	Remark	Concept pairs	Remark
Hypo(ህ ግ , ፍ ት ህ)	√	Hypo(ት ም ር ት ,መ ያ)	√	Hypo(ስ ራ ,ከ ል ል)	×
Hypo(ፓ ር ላ ማ,ህ ግ)	√	Hypo(መ ን ግ ስ ት ,ፍ ት ህ)	√	Hypo(ህ ግ ,ኢ ት ዮ ጵ ያ)	×
Hypo(ህ ግ ,መ ሰ ረ ት)	√	Hypo(ኢ ት ዮ ጵ ያ ,አ ገ ር)	√	Hypo(አ ገ ር ,አ ለ ም)	×
Hypo(ህ ግ ,አ ሰ ራ ር)	√	Hypo(ከ ል ል ,ቢ ሮ)	√	Hypo(ከ ል ል ,ስ ራ)	×
Hypo(ህ ግ ,መ ን ግ ስ ት)	√	Hypo(ድ ር ጅ ት ,ፕ ሮ ጀ ክ ት)	√	Hypo(ሱ ዳ ን ,ደ ቡ ብ)	×
Hypo(ተ ማ ሪ ,መ ያ)	√	Hypo(ኢ ት ዮ ጵ ያ ,መ ን ግ ስ ት)	√	Hypo(መ ራ ር ,ከ ል ል)	×
Hypo(ቻ ይ ና ,አ ገ ር)	√	Hypo(አ ማ ራ ,ከ ል ል)	√	Hypo(ወ ረ ዳ ,ከ ል ል)	×
Hypo(አ ማ ራ ,ከ ል ል)	√	Hypo(ወ ረ ዳ ,ቀ በ ሌ)	√	Hypo(ሱ ዳ ን ,ግ ብ ጽ)	×
Hypo(ብ ሄ ር ,ቋ ን ቋ)	√	Hypo(መ ን ግ ስ ት ,ህ ግ)	√	Hypo(ከ ል ል ,ል ማ ት)	×
Hypo(ከ ል ል ,ማ እ ከ ል)	√	Hypo(ብ ሄ ር ,ማ ን ነ ት)	√	Hypo(ህ ይ ማ ኖ ት ,ኢ ት ዮ ጵ ያ)	×
Hypo(ድ ር ጅ ት ,ስ ራ)	√	Hypo(ህ ድ ይ ,ዞ ን)	√	Hypo(ኢ ት ዮ ጵ ያ ,አ ፍ ሪ ካ)	×
Hypo(ኤ ር ት ራ ,ጦ ር)	√	Hypo(አ ፍ ሪ ካ ,አ ህ ጉ ር)	√	Hypo(አ ገ ር ,አ ፍ ሪ ካ)	×
Hypo(ብ ሄ ር ,ማ ን ነ ት)	√	Hypo(አ ገ ር ,ህ ዝ ብ)	√	Hypo(አ ማ ራ ,ት ግ ራ ይ)	×
Hypo(ዞ ን ,ከ ል ል)	√	Hypo(ነ ጻ ነ ት ,ህ ይ ማ ኖ ት)	√	Hypo(አ ውሮ ,ፓ ር ላ ማ)	×
Hypo(ል ማ ት ,ፕ ሮ ጀ ክ ት)	√	Hypo(አ ፍ ሪ ካ ,አ ገ ር)	√	Hypo(ግ ብ ጽ ,አ ፍ ሪ ካ)	×
Hypo(ግ ብ ር ና ,ስ ራ)	√	Hypo(ድ ር ጅ ት ,መ ስ ተ ዳ ድ ር)	√	Hypo(ኢ ት ዮ ጵ ያ ,አ ለ ም)	×
Hypo(ድ ር ጅ ት ,ል ማ ት)	√	Hypo(ግ ብ ር ና ,ፕ ሮ ጀ ክ ት)	√		
Hypo(ህ ድ ያ ,ዞ ን)	√	Hypo(ስ ራ ,መ ታ ቀ ድ)	√		
Hypo(ፖ ለ ቲ ካ ,ም ህ ዳ ር)	√	Hypo(ግ ን ኙ ነ ት ,አ ዲ ስ)	×		

Table 5.7. Extracted Concepts with Part-of Relation Represented by FOPC

Concept pairs	Remark	Concept pairs	Remark	Concept pairs	Remarks
Part(ህ ግ ,ፍ ት ህ)	√	Part(ል ማ ት ,ስ ራ)	√	Part(ፓ ር ቲ ,ነ ጻ ነ ት)	√
Part(ዞ ን ,ከ ል ል)	√	Part(አ ገ ር ,አ ለ ም)	√	Part(ድ ር ጅ ት ,ግ ብ ር)	√
Part(ስ ራ ,ከ ል ል)	√	Part(አ ገ ር ,ሱ ዳ ን)	√	Part(ድ ር ጅ ት ,ፕ ሮ ጀ ክ ት)	√
Part(ህ ግ ,ፓ ር ላ ማ)	√	Part(ወ ረ ዳ ,አ ስ ተ ባ ባ ሪ)	√	Part(ል ማ ት ,ት ግ ራ ይ)	√
Part(ህ ግ ,መ ሰ ረ ት)	√	Part(ወ ረ ዳ ,ቀ በ ሌ)	√	Part(አ ለ ም ,አ ገ ር)	√
Part(ህ ግ ,አ ሰ ራ ር)	√	Part(ከ ል ል ,ቢ ሮ)	√	Part(ል ማ ት ,መ ሰ ረ ት)	√
Part(ዞ ን ,ቀ በ ሌ)	√	Part(ወ ረ ዳ ,ከ ል ል)	√	Part(ግ ብ ጽ ,አ ፍ ሪ ካ)	√
Part(ዞ ን ,አ ስ ተ ባ ባ ሪ)	√	Part(ድ ር ጅ ት ,ስ ራ)	√	Part(ል ማ ት ,ድ ር ጅ ት)	√
Part(ህ ግ ,ኢ ት ዮ ጵ ያ)	√	Part(ወ ረ ዳ ,ህ ድ ይ)	√	Part(ግ ብ ር ,ፕ ሮ ጀ ክ ት)	√
Part(ህ ግ ,መ ን ግ ስ ት)	√	Part(ል ማ ት ,ከ ል ል)	√	Part(ል ማ ት ,ፕ ሮ ጀ ክ ት)	√
Part(ስ ራ ,ህ ብ ረ ተ ሰ ብ)	√	Part(መ ን ግ ስ ት ,ነ ጻ ነ ት)	√	Part(አ ገ ር ,መ ን ግ ስ ት)	√

Part(ልማት,ስራ)	√	Part(ከልል,ልማት)	√	Part(ልማት,ሀብረተሰብ)	√
Part(ከልል,ፕሮጀክት)	√	Part(ገጠር,ከልል)	√	Part(ከልል,መሰተዳድር)	√
Part(ተማሪ,መያ)	√	Part(ፓርቲ,ተቃዋሚ)	√	Part(ግብርና,ስራ)	√
Part(ድር ጅት,ከልል)	√	Part(ኢትዮጵያ,መን ግስት)	√	Part(ኤርትራ,ጦር)	√
Part(አፍሪካ,አገር)	√	Part(አገር,አፍሪካ)	√	Part(አመልድ,አማራ)	√
Part(ምህዳር,ፖለቲካ)	√	Part(ስፖርት,አትሌቲክስ)	√	Part(መሰረት,ፍትህ)	√
Part(ነጻነት,ሀይማኖት)	√	Part(መን ግስት,ፍትህ)	√	Part(ህብረት,አገር)	√
Part(ፓርላማ,መን ግስት)	√	Part(ኢትዮጵያ,አለም)	√	Part(መን ግስት,መሰረት)	√
Part(አድገት,ሀብረተሰብ)	√	Part(መን ግስት,ኢፌዴሪ)	√	Part(መን ግስት,ኢትዮጵያ)	√
Part(ድር ጅት,መን ግስታዊ)	√	Part(ኢትዮጵያ,ኢፌዴሪ)	√	Part(ሀብረተሰብ,ከተማ)	√
Part(ኢኮኖሚ,ሀብረተሰብ)	√	Part(ኢትዮጵያ,አፍሪካ)	√	Part(ገብረስላሴ,ታዋቂ)	√
Part(ድር ጅት,መሰተዳድር)	√	Part(ትምህርት,ቴክኒክ)	√	Part(አገር,ግንኙነት)	×
Part(ቢሮ,ስራ)	×	Part(ከልል,ማእከል)	×	Part(አገር,ኢትዮጵያ)	×
Part(ስራ,ደረጃ)	×	Part(ደረጃ,መታቀድ)	×	Part(አገር,አንቅስቃሴ)	×
Part(ቢሮ,ልማት)	×	Part(አገር,ጉብኝት)	×	Part(ፖለቲካ,ምርር)	×
Part(ስራ,ጀምር)	×	Part(አቀፍ,ኢትዮጵያ)	×	Part(ኢትዮጵያ,ፍትህ)	×
Part(ከልል,ስራ)	×	Part(መን ግስት,አወር)	×	Part(ኢትዮጵያ,ግንኙነት)	×
Part(መን ግስት,አመት)	×	Part(ኢትዮጵያ,ግብጽ)	×	Part(ኢትዮጵያ,አገር)	×
Part(መለስ,አመት)	×	Part(ኢትዮጵያ,አዲስ)	×	Part(ግንኙነት,አዲስ)	×
Part(ሱድን,ደቡብ)	×	Part(ኢትዮጵያ,አመት)	×	Part(መን ግስት,ሀይማኖት)	×
Part(አቀፍ,አገር)	×	Part(ከልል,ጀምር)	×	Part(ከልል,ደረዳ)	×

In Tables 5.6 and 5.7, the relation is shown by using first order predicate calculus (FOPC) in which the two arguments called concepts are related by the predicate which is currently denoted as either "hypo" or "part". For example in Hypo (ኢትዮጵያ,አገር) and Part(ዞን,ከልል), the FOPC shows that ኢትዮጵያ (Ethiopia) is a concept which is type of concept of አገር (country) and ዞን (zone) is a concept which is part of the concept of ከልል (Region) respectively. Consider the following Amharic sentence, አመልድ መን ግስታዊ የሆነ ድርጅት ሲሆን ዋና መስሪያ ቤቱ በአማራ ከልል ባህርዳር ከተማ ይገኛል።, to be represented as a semantic network using the knowledge base we built and shown on Tables 5.6 and 5.7. It is represented using semantic network in Figure 5.2.

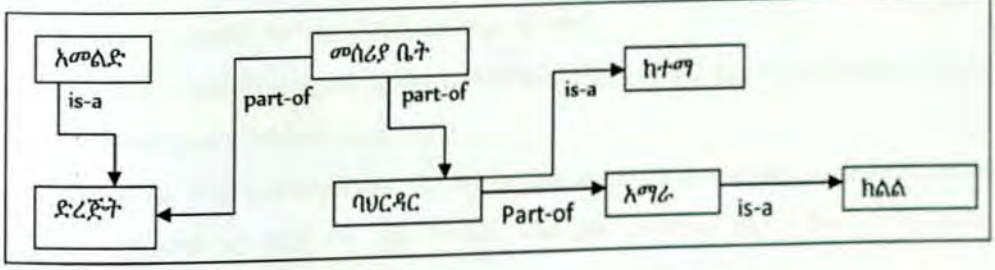


Figure 5.2. Example Semantic Network for Amharic Sentence

The accuracy of the system to extract pairs of concepts with both type-of and part-of relations are calculated using the results of the third experiment. Concepts in this case are synsets that are composed of synonym terms. These synsets are identified by unique identification number. The accuracy of the system to extract both type-of and part-of relations between pairs of concepts is shown in Table 5.8.

Table 5.8. Evaluation of the System

Type-of relation			Part-of relation		
Total extracted	Correct	Accuracy	Total extracted	Correct	Accuracy
54	37	68.5%	99	71	71.7%

In general, the three stage experimentation shows that when size of corpus increases the frequency of terms increased and co-occurrence analysis becomes smooth. Among the 4325 documents in the corpus only 1065 are tagged with part of speech manually. So, whenever untagged document is added to the corpus, the probability that untagged terms in the corpus will increase. This leads to inability to extract enough number of noun concepts for our research. This is the reason behind not to increase the number of documents in the corpus beyond we expressed above. The accuracy of the system to extract the type-of and part-of relations between concepts (synsets) from free text corpus is 68.5% and 71.7% respectively.

Because our system is depending on the available of POS tagger, we have some small tagged data for the experiment. Even if the main target of our research is to build a knowledge base as a form of semantic network in a short amount of time trading off some precision lose for the speed of construction, we have got some (probably major) obstacles we faced and are a means to decrease the precision of the system. Some of the problems were:

- There is no well developed and implemented Amharic part of speech tagger to categorise concepts as nouns, verbs, adjectives and adverbs.
- There is no well developed Amharic stemmer that is used to stem different inflections of a word into one common form.
- There is no well implemented Amharic named entity recognizer to extract entities like proper names so that we can detect relations between them easily. Named entity recognizer is very crucial tool for information extraction.

CHAPTER SIX
CONCLUSION AND RECOMMENDATION

This chapter presents the conclusion and recommendations for future study based on our findings.

6.1 Conclusion

A semantic network is a directed graph consisting of vertices, representing concepts, and edges representing type of relations. It is often used as a form of knowledge representation. Semantic networks have many benefits for machine translation, query optimization, document classification, language teaching and translation, and information retrieval. There are a number of researches done so far related to automatic construction of semantic networks. Those works implement different approaches to extract concepts and their relations automatically from free text and other resources. Knowledge-based, corpus- based and hybrid are the main approaches used in many researches.

The first approach is knowledge -based in which construction of semantic networks based on the creation of semantic paths between words in a text using the thesaurus, like WordNet. Early approaches in this field used gloss words from the respective word definitions in order to build semantic networks from text. The expansion of WordNet with semantic relations that cross parts of speech has added more possibilities of semantic network construction from text.

Corpus-based approach tries to identify the degree of similarity between words using information exclusively derived from large corpora. Two methods considered, namely: pointwise mutual information (PWMI) and latent semantic analysis (LSA).

Hybrid approach is the third approach which combines both the hierarchy of the used thesaurus (knowledge-based), and statistical information from large corpora (corpus-based). ASNet is based on the hybrid approach by utilizing the available semantic relations from Amharic WordNet synsets and acquiring more statistical information from the corpus.

WordSpace model is a special form of vector space model in which the general idea behind WordSpace model is to use distributional statistics to generate high-dimensional vector spaces, in which words are represented by context vectors whose relative directions are assumed to indicate semantic similarity. This assumption is motivated by the distributional hypothesis, which states that words with similar meanings tend to occur in similar contexts. The core idea behind WordSpace model is that words and concepts are represented by points in a mathematical space, and this representation is learned from text in such a way that concepts with similar or related meanings are near to one another in that space. Such models are designed to represent words and documents in terms of underlying concepts, and as such can be used for many semantic matching tasks such as knowledge representation and concept matching.

A major effort was made in identifying and defining a formal set of steps for automatic Construction of semantic network of Amharic noun concepts from free text corpus. The construction model of our semantic network involves the creation of index file for the collected news text corpus, development of WordSpace based on the index file, searching the WordSpace to generate semantically related concepts for a given Amharic WordNet term, generate patterns for a specific relation using entries of Amharic WordNet and detect relations between each pair of concepts among the related concepts using those patterns. The design of the system was made and implemented using java programming language. We have made a test using Amharic WordNet and free text corpus, from Ethiopian news agency, and Walta information Center for each semantic relations.

Finally, the effectiveness of the proposed system was demonstrated to check the accuracy. The Three step experimentation was done on different datasets and we have got reasonable results from each experiment. The results were traced manually to check whether it is accurate, contextually, or not.

All the papers that we have reviewed in this work are done for English language, which is resourceful and in different aspects. Most of the systems used different tools as component for full implementation. We cannot compare our work with these systems because of the following key points:

- The morphological structure of Amharic language is more inflectional and is difficult to analyaze Amharic texts easily as English one. There is no well designed and implemented

stemmer for Amharic language that changes all the inflections of a word in to one common form.

- There are no enough resources and tools that support our research like those works done for English language. Different NLP tools like POST matters for the best result of the experiment.
- We have done our work with minimum supervision by providing simple seed terms as input so that the system can learn from these inputs and generate approparte outputs.
- Most of the works we have reviewed in this paper are evaluating simply by counting how many concepts as nodes and relations as edges that the network is composed of. So they left out the idea of accuracy of their systems becuase of the complexity of the constructed semantic network.
- We have shown each and every intermidate results along with the average accuracy of the system to extract both the type-of and part-of relations from the free text corpus.
- Even if we have used the intevening words pattern to extract realtions between given concepts, there is no as such common text patterns as compared to text patterns in English language.

The results shows that, using some additional NLP tools like parser, named entity recognizer, POS tagger, good stemmer and multi term extractor can lead to develop full fledged Amharic Semantic Network (ASNet) that can learn automatically from free text corpora or on the web.

6.2 Recommendation

Semantic network is an important knowledge representation technique for different NLP applications and the approach we followed to build semantic network of Amharic concepts is expected to be a significant contribution to the field, and mainly to researchers working on various aspects of the Amharic language, such as in Amharic search engine, document classification, question answering machines ...etc. Therefore, the researchers in the area can use the design of our model or the implemented system for knowledge representation as a component in their research.

As a future work, we would like to suggest the following key points:

- Amharic concepts are sometimes represented with multi-word terms. So extracting multi-word terms from corpus automatically and store them as index term is mandatory.

- We have used manually developed Amharic WordNet for concept, pattern and relation extraction processes. No full fledged Amharic WordNet is available and constructing it manually is tedious. So constructing such knowledge base like ASNet in unsupervised way with no seed inputs is important and time efficient.
- Other than these lexical relations, meronym (part-of), hierarchy (type-of), synonym and antonym, there are other non-taxonomic relations between concepts. Extracting all possible relations between each pair of concepts from free Amharic language documents is also very valuable.
- Even if most of the concepts are found on nouns, sometimes we can get concepts on verbs and adjectives. So, incorporating additional methods to extract relations from other word category like that of adjectives and adverbs is also essential for full fledged knowledge base.

REFERENCES

[1]. Amaryllis, D., Robert, A. (1979), Logic and Semantic Networks, *ACM* , 9.

[2]. Berners-Lee, T. (2001), The Semantic Web, *Scientific American* , 34-43.

[3]. Bos, J., Clark, S., Steedman, M., Curran, J. R., and Hockenmaier, J. (2004), Wide-coverage semantic representations from a CCG parser, *In Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, (S. 1240–1246).

[4]. Brian, S. (2007), ASKNet: Automated Semantic Knowledge Network, *Oxford University Computing Laboratory* , 6.

[5]. Brickley, D. (2004), RDF Vocabulary Description Language 1.0: RDF Schema, W3C.

[6]. C. Gellbaum, e. (1992), WordNet:An Electronic Lexical Database, *MIT press* , 87.

[7]. Clark, S., James, R. (2004), Parsing the WSJ using CCG and log-linear models, *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, (S. 104-111).

[8]. Cochran, G. (1977), Sampling Techniques, New York, U.S.A: Wiley Publishers.

[9]. Domingos, P., Matthew, R. (2006), Markov logic networks, Machine Learning. 62.

[10]. Dominic, W., Trevor, C. (2010), The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics, *Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010)* (S. 7).

[11]. Doug, D., Oren, E., Stephen, S., and Daniel, S. W. (2008), Learning Text Patterns for Web Information Extraction and Assessment, *University of Washington* , 6.

[12]. George, T. (2010), Text Relatedness Based on a Word Thesaurus, *Journal of Artificial Intelligence Research* , 39.

[13]. Landauer, T., Foltz, W. P., Laham, D. (1998), Introduction to latent semantc analysis, *Discourse Processes* , 25.

[14]. Liu, H., and Singh, P. (2004), ConceptNet, *BT Technology Journal* , 16.

[15]. Lucene. (2012), *Lucene Tutorial*, Abgerufen am 17. 05 2012 von Lucene Solr Elastic Search consultant: http://www.lucenetutorial.com

[16]. Michele, B., Michael, J. C., Stephen, S., Matt, B., and Oren, E. (2007), Open information extraction from the web, *In Proc. IJCAI-2007, Hyderabad* , 7.

[17]. Marvin, L., Bender, H. W. (1976), The Ethiopian Writing System, *London, Oxford University press* .

[18]. MEKONNEN, A. (2009), *AUTOMATIC THESAURUS CONSTRUCTION FOR AMHARIC TEXT RETRIEVA,* Addis Ababa, unpublished.

[19]. Miller, G. A. (1995), WordNet: A Lexical Database for English, *COMMUNICATIONS OF THE ACM* , 3.

[20]. Mostafa, H. (2007), *N-Gram Pattern*, Abgerufen am 20. 06 2012 von Code Project: http://www.codeproject.com/Articles/20423/N-gram-and-Fast-Pattern-Extraction-Algorithm

[21]. Navigli, R. (2008), A structural approach to the automatic adjudication of word sense disagree- ments, Natural Language Engineering, 547-579.

[22]. Nega, A., and Willet, P. (2002), Stemming of Amharic Words for Information Retrieval, *In Literary and Linguistic Computing, Oxford, Oxford University press* , 1-17.

[23]. Oren, E., Michael, C., and Doug, D. (2004), Web-Scale Information Extraction in KnowItAll, *ACM* , 11.

[24]. Ravichandran, D., Hovy, E. (2002), Learning surface text patterns for a question answering system, *Proc 40th ACL Conf*, (S. 6).

[25]. Richardson, D., William, B., and Lucy, V. (1998), MindNet: Acquiring and structuring semantic information from text, *In Proceedings of COLING '98*, (S. 5).

[26]. Sahlgren, M. (2006), *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector space,* Stockholm

[27]. Seamons, K. (1997), Document Ranking and the Vector Space Model, *IEEE* , 9.

[28]. Smith, A., Smith ,E., Centre, K., and Factors, H. (2003), Automatic Extraction of Semantic Networks from Text using Leximancer, *Proceedings of HLT-NAACL* , 2.

[29]. Stanley, K., Pedro, D. (2010), Extracting Semantic Networks from Text Via Relational Clustering, *In Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases-Part I*, (S. 16).

[30]. Steedman, M. (2000), The Syntactic Process, Cambridge, The MIT Press.

[31]. Tessema, M., Meron, S., and Teshome, k. (2010), The Need for Amharic WordNet, *IEEE* , 4.

[32]. Turney, P. D. (2006), Similarity of Semantic Relations, *Association for Computational Linguistics* , 38.

[33]. Vossen, P. (1998), EuroWordNet: A multilingual database with lexical semantic networks, *Kluwer Academic Publishers, Dordrecht* , 3.

[34]. Xiaoli, Z. F., and Carla, B. E. (2003), Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach, *School of Electrical and Computer Engineering, Purdue University* , 8.

[35]. Filip, H. (2008), Introduction to Natural Language Semantics: University of Florida.

[36]. Nazihah, A., Alan, D., Giorgos, L. (2011), Spreading Activation for Web Scale Reasoning: Promise and Problems, *ACM.*

APPENDICES

Appendix I:

List of Stop Words

ሀ ያ	ሰ ባ ት	በ ር ካ ታ	ና ቸ ው	እ ር ስ ዋ	ከ ዛ	ይ ሆ ና ሉ
ሀ ምሳ	ሰ አ ት	በ ቀ ር	እ ሁ ን	እ ር ሷ	ወ ዘ ተ	ይ ሆ ና ል
ሁ ለ ት	ሲ ሆ ን	በ ኩ ል	እ ለ	እ ባ ክ ህ	ወ ይ ም	ደ ግ ሞ
ሁ ሉ	ሲ ል	በ ውጭ	አ ምስ ት	እ ነ ር ሱ	ወ ይ ስ	ዳ ሩ
ሆ ነ	ሳ ለ	በ ጣ ም	አ ስ ራ	እ ና	ወ ደ	ድ ረ ስ
ሆ ን	ስ ለ	በ ፊ ት	አ ስ ር	እ ና	ወ ደ ፊ ት	ጊ ዜ
ሆ ይ	ስ ል ሳ	ባ ለ	አ ራ ት	እ ና ንተ	ወ ዲ ህ	ጋ ር
ኋ ላ	ስ ም ን ት	ብቻ	አ ር ባ	እ ኔ	ወ ዴ ት	ግ ን
ላ ይ	ስ ነ	ብ ዙ	አ ቤ ቱ	እ ን	ዋ ና	ጥ ቂ ት
ሌ ላ	ስ ን ት	ተ ለ ያ የ	አ ን ተ	እ ን ኪ ያ ስ	ወ ስ ጥ	ጸ ረ
መ ሆ	ስ ን ኳ	ተ ጨ ማ ሪ	አ ን ቺ	እ ን ኳ	ወ ጪ	እ ነ ሆ
ማ ን	ስ ድ ሳ	ታ ች	አ ን ድ	እ ን ደ	ዘ ን ድ	እ ን ዲ ህ
ም ን	ስ ድ ስ ት	ታ ች	አ ን ድ	እ ን ደ ገ ና	ዘ ጠ ና	
ም ን ድ ር	ሶ ስ ት	ት ና ን ት	አ ይ ደ ለ ም	እ ን ዲ ሁ ም	ዘ ጠ ኝ	
ም ክ ን ያ ቱ ም	ራ ሱ	ነ ህ	አ ል ፍ	እ ን ጂ	ዚ ህ	
ም ክ ን ያ ት	ሺ ህ	ነ ሽ	አ ሱ	ሆ ን	ዚ ህ	
ሰ ላ ሳ	ቀ ጥ ሎ	ነ በ ር	አ ስ ከ	እ ን ግ ዲ ህ	ያ ህ ል	
ሰ ማ ን ያ	ቁ ጥ ር	ነ ው	እ ሷ	እ ኛ	ይ ሁ ን	
ሰ ሞ ን	በ ላ ይ	ነ ገ ር	እ ር ሱ	እ ያ ን ዳ ን ዱ	ይ ህ	
ሰ ባ	በ ር ስ	ና ት	እ ር ስ	እ ጅ ግ	ይ ህ ን	

Appendix II:

Seeds

List of Type-of Relation Seed Pairs for Experiment 3

ታጋይ:ሰው	ማሳ:ሃብት	ስራ:ዝግጅት	ጤና:ምግብ	ኩሬ:ውሃ	ሰው:አገር	ገቢ:ውጤት
ፕሮግራም:መምሪያ	ማምረቻ:ሃብት	ማሳ:እጸዋት	ሰው:ህብረተሰብ	ጤና:አሴት	ኑሮ:ጤና	መሪ:ቡድን
ገቢ:ሃብት	ሸቀጥ:ሙያ	ሸቀጥ:ሱቅ	ምግብ:እህል	ማሳ:ገቢ	ምግብ:ማሳ	ሃይል:እቃ
መሪ:ህዘብ	ሶማሌ:ክልል	ሸቀጥ:ስራ	ደርግ:ፓርቲ	መረጃ:ወሬ	ጅጅጋ:ቦታ	ፕላን:ጉዳይ
ነጋዴ:ሰው	ስጦታ:አሴት	በጀት:ፕላን	ባንክ:ሃብት	ጀርም:ደዌ	በጀት:እቅም	በጀት:ገንዘብ
ምግብ:ጤና	ችግር:ተረጅ	ሙከራ:ሙያ	ስልት:ፈጠራ	አማራ:ክልል	በጀት:ሃብት	ኮሌጅ:ተቋም
ገበሬ:ገጠር	እጸዋት:ደን	ልማት:አሴት	ኦሮሚያ:ክልል	ብሄር:ክልል	ሃብት:ለጨጥ	ገጠር:አፍሪካ
ጋዘጣ:ፕሬስ	መግለጫ:ዘገባ	ሰራተኛ:ሰው	አፍሪካ:ክልል	ሸቀጥ:ሃብት	ስጦታ:እርዳታ	አህዴድ:ቡድን
ኮሌጅ:ተቋም	ኢህዴግ:ፓርቲ	ኩባንያ:ሃብት	ትኩረት:ቡድን	በጀት:ሃብት	አቅም:ሃብት	ማምረቻ:ካምፓኒ
ችግር:ጠብ	ትኩረት:እቅድ	ፕሮግራም:ጉዳይ	ኩባንያ:ካምፓኒ	ልምድ:ስልት	ጋምቤላ:ክልል	
ጤና:ሰላም	ኢፌድሪ:አገር	ቤንሻንጉል:ክልል	ህክምና:ተቋም	አሜሪካ:አገር	ኢህዴግ:ቡድን	
ኮሌጅ:ተቋም	ማዳበሪያ:ኬሚካል	ሰራዊት:ጠባቂ	ቤተክርስትያን:ወግ	ኢፌድሪ:አገር	ፌደራል:ፓርቲ	
ልምድ:ፈጠራ	ገንዘብ:ሃብት	እድገት:ሃብት	ሪሰርች:ስራ	ፕሮግራም:ዝግጅት	ሪሰርች:ፕሮግራም	

List of Part-of Relation Seed Pairs for Experiment 3

ገ ቢ:ጥ ረ ት	ገ ቢ:ስ ራ	ቀ ን :ወ ር	ኩ ሬ :ቦ ታ	ወ ር :አ ሙ ት	ሰ ው:ቋ ን ቋ
ጤ ና :ም ግ ብ	ገ ቢ:ው ጤ ት	ጤ ና :አ ሴ ት	ስ ራ :አ ሴ ት	ጤ ና :ሰ ላ ም	ዞ ን :ክ ል ል
ወ ር :አ ሙ ት	ስ ራ :ሪ ሰ ር ች	መ ሪ :ሙ ያ	ዲ ግ ሪ :ሙ ያ	መ ሪ :ኢ ፌ ድ ሪ	ጤ ና :ህ ክ ም ና
ህ ግ :መ ን ግ ስ ት	ዞ ን :ክ ል ል	መ ሪ :አ ገ ር	ባ ህ ል :ሰ ው	ሸ ቀ ጥ :ሱ ቅ	ች ሎ ት :ህ ግ
ሸ ቀ ጥ :ስ ራ	ክ ተ ማ:ዞ ን	ም ግ ብ :ካ ፌ	አ ገ ር :አ ለ ም	ሃ ብ ት :ስ ራ	ክ ተ ማ:ወ ር ዳ
ተ ማ ር :ኮ ሌ ጅ	ወ ጤ ት :ስ ራ	ጅ ጅ ጋ :ቦ ታ	ጅ ጅ ጋ :ሶ ማ ሌ	ባ ህ ል :ክ ል ል	ፓ ር ቲ :ክ ል ል
ገ በ ሬ :ገ ጠ ር	ዲ ግ ሪ :ኮ ሌ ጅ	ቀ በ ሌ :ወ ር ዳ	ተ ረ ጅ :አ ፍ ሪ ካ	ቁ ጠ ባ :ባ ን ክ	ባ ህ ል :ብ ሄ ር
ባ ን ክ :ሃ ብ ት	ሸ ቀ ጥ :ክ ፍ ል	ብ ሄ ር :አ ገ ር	ቀ በ ሌ :ወ ረ ዳ	ሸ ቀ ጥ :ሃ ብ ት	በ ጀ ት :አ ገ ር
በ ጀ ት :ፓ ር ቲ	ጉ ባ ኤ :አ ማ ራ	ክ ብ ር :ህ ዘ ብ	ክ ፍ ያ :ሃ ብ ት	ጋ ዘ ጣ :ፕ ሬ ስ	ቡ ድ ን :አ ገ ር
ክ ተ ማ:አ ገ ር	ቡ ድ ን :ፓ ር ቲ	ቋ ን ቋ :ብ ሄ ር	ገ ጠ ር :አ ፍ ሪ ካ	ሃ ይ ል :ም ግ ብ	ደ ር ግ :ኢ ፌ ድ ሪ
አ ማ ራ :ኢ ፌ ድ ሪ	ሃ ይ ል :ሃ ብ ት	ፓ ር ቲ :አ ገ ር	አ ፍ ሪ ካ :አ ለ ም	ህ ረ ር :ኢ ፌ ድ ሪ	ሶ ማ ሌ :ኢ ፌ ድ ሪ
ተ ረ ጅ :ኢ ፌ ድ ሪ	በ ጀ ት :ኢ ፌ ድ ሪ	ባ ህ ል :ኢ ፌ ድ ሪ	አ ድ ግ ት :ሃ ብ ት	ባህል:ህብረተሰ	ዘ መ ድ :ህ ብ ረ ተ ሰ ብ
ወ ር ዳ :ዞ ን	ብ ሄ ር :ኢ ፌ ድ ሪ	ዲ ግ ሪ :መ ም ህ ር	ሪ ሰ ር ች :ካ ም ፓ ኒ	መ ሬ ት :ቦ ታ	አ ገ ር :አ ለ ም
ክ ተ ማ:አ ገ ር	ወ ረ ዳ :ዞ ን	አ ቅ ም:ቦ ታ	ኢ ኮ ኖ ሚ ኢ ት ዮ ጵ ያ	አ ገ ር :ኔ ፓ ድ	መ ካ ራ :ኮ ሌ ጅ
ቅ ሪ ት :ክ ል ል	መ ለ የ :ጉ ዳ ይ	ህ ዝ ብ :አ ገ ር	ኢ ት ዮ ጵ ያ :ኔ ፓ ድ	ት ፎ ዞ :ቡ ድ ን	ሃ ብ ት :አ ገ ር
ፖ ሊ ስ :ህ ዝ ብ	ም ሁ ር :ኮ ሌ ጅ	ገ ጠ ር :ክ ል ል	ህ ብ ረ ተ ሰ ብ :ክ ል ል	አ ቅ ም:ሃ ብ ት	ል ም ድ :ብ ሄ ር
ክ ል ል :አ ገ ር	አ ገ ር :አ ለ ም	ቅ ሪ ት :አ ገ ር	ት ግ ራ ይ :ኢ ት ዮ ጵ ያ	ክ ል ል :ኢ ት ዮ ጵ ያ	ቤ ን ሻ ን ጉ ል :ኢ ፌ ድ ሪ
ቅ ሪ ት :ኢ ፌ ድ ሪ	አ መ ት :ዘ መ ን	ጎ ዳ ና :ሃ ብ ት	ሪ ሰ ር ች :ፕ ሮ ግ ራ ም	መ ም ህ ር :ኮ ሌ ጅ	ዲ ሞ ክ ራ ሲ :መ ን ግ ስ ት
ተ ማ ር :ዩ ኒ ቨ ር ሲ ቲ	ገ ጠ ር :ኢ ፌ ድ ሪ	ቅ ሪ ት :አ ፍ ሪ ካ	ማ ም ረ ቻ :ሃ ብ ት	አ ወ ሮ ፓ :አ ገ ር	ብ ሄ ረ ሰ ብ :ኢ ት ዮ ጵ ያ
ድ ን በ ር :አ ገ ር	ባ ለ መ ያ :መ ያ	መ ም ሪ ያ :ፈ ይ ል	ድ ር ጅ ት :ኢ ፌ ድ ሪ	አ ማ ራ :ኢ ት ዮ ጵ ያ	ፕ ሮ ጀ ክ ት :ፕ ሮ ግ ራ ም
አ ሮ ሚ ያ :ኢ ፌ ድ ሪ	ኩ ባ ን ያ :ሃ ብ ት	ሰ ራ ዊ ት :ጠ ባ ቂ	ድ ን በ ር :ክ ል ል	ል ማ ት :ሃ ብ ት	መ ን ግ ስ ት :አ ገ ር
ካ ም ፓ ኒ :ፓ ር ቲ	ድ ንበር:ኢ ፌ ድ ሪ	ጋ ምቤላ:ኢ ፌ ድ ሪ	መ ን ገ ድ :ፕ ሮ ጀ ክ ት	ኢ ትዮጵያ:እፍሪከ	መ ን ግ ስ ት :ኔ ፓ ድ
መ ን ግ ስ ት :አ ማ ራ	ኒ ዮ ር ክ:አ ሜ ሪ ካ	ኢ ፌ ድ ሪ:አ ፍ ሪ ካ	መ ን ግ ስ ት:ኢ ት ዮ ጵ ያ	ህ ብ ረ ተ ሰ ብ:አ ገ ር	ህ ብ ረ ተ ሰ ብ:ቋ ን ቋ

Declaration

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

Declared by:

Name: Alelegn Tefera

Signature: _____

Date: Apr, 2, 2013

Confirmed by advisor:

Name: Yaregal Assabie (PhD)

Signature: _____

Date: April 02, 2013