Automatic Construction of Amharic Semantic Networks From Unstructured Text Using Amharic WordNet

Alelgn Tefera
Department of Computer Science
Jigjiga University, Ethiopia
alelgn.tefera@gmail.com

Yaregal Assabie
Department of Computer Science
Addis Ababa University, Ethiopia
yaregal.assabie@aau.edu.et

Abstract

Semantic networks have become key components in many natural language processing applications. This paper presents an automatic construction of Amharic semantic networks using Amharic WordNet as initial knowledge base where intervening word patterns between pairs of concepts in the WordNet are extracted for a specific relation from a given 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 use the WordSpace model for extraction of semantically related concepts and relation identification among these concepts utilizes the extracted text patterns. The system is designed to extract "part-of" and "type-of" relations between concepts which are very popular and frequently found between concepts in any corpus. The system was tested in three phases with text corpus collected from news outlets, and experimental results are reported.

1 Introduction

A semantic network is a network which represents semantic relations among concepts and it is often used to represent knowledge. A semantic network is used when one has knowledge that is best understood as a set of concepts that are related to one another. Concepts are the abstract representations of the meaning of terms. A term can be physically represented by a word, phrase, sentence, paragraph, or document. The relations between concepts that are most commonly used in semantic networks are *synonym* (similar concepts), *antonym* (opposite concepts), *meronym/holonym* ("part-of" relation between concepts), and *hyponym/hypernym* ("type-of" relation between concepts). Knowledge stored as semantic networks can be represented in the form of graphs (directed or undirected) using concepts as nodes and semantic relations as labeled edges (Fellbaum, 1998; Steyvers and Tenenbaum, 2005). 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 processing (NLP) 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 for document classification using the knowledge stored in it (Berners-Lee, 2001; Sahlgren, 2006; Smith, 2003).

Approaches commonly used to automatically construct semantic networks are knowledge-based, corpus-based and hybrid approaches. In the knowledge-based approach, relations between two concepts are extracted using a thesaurus in a supervised manner whereas corpus-based approach extracts concepts from a large amount of text in a semi-supervised method. Hybrid approach combines both the hierarchy of the thesaurus and statistical information for concepts measured in large corpora (Dominic and Trevor, 2010; George *et al*, 2010; Sahlgren, 2006). Over the past years, several attempts have been made to develop semantic networks. Among the widely known are ASKNet (Harrington and Clark, 2007), MindNet (Richardson *et al*, 1998), and Leximancer (Smith, 2003). Most of the semantic networks constructed so far assume English text

as corpus. However, to our best knowledge, there is no system that automatically constructs semantic networks from unstructured Amharic text.

This paper presents an automatic construction of semantic networks from unconstrained and unstructured Amharic text. The remaining part of this paper is organized as follows. Section 2 presents Amharic language with emphasis to its morphological features. The design of Amharic semantic network construction is discussed in Section 3. Experimental results are presented in Section 4, and conclusion and future works are highlighted in Section 5. References are provided at the end.

2 Amharic Language

Amharic is a Semitic language spoken predominantly in Ethiopia. It is the working language of the country having a population of over 90 million at present. The language is spoken as a monther tongue by a large segment of the population in the northern and central regions of Ethiopia and as a second language by many others. It is the second most spoken Semitic language in the world next to Arabic and the most commonly learned second language throughout Ethiopia (Lewis *et al*, 2013). Amharic is written using a script known as *fidel* having 33 consonants (basic characters) out of which six other characters representing combinations of vowels and consonants are derived for each character.

Derivation and inflection of words in Amharic is a very complex process (Amare, 2010; Yimam, 2000). Amharic nouns and adjectives are inflected for number, gender, definitnesss, and cases. On the other hand, Amharic nouns can be derived from:

- *verbal roots* by infixing various patterns of vowels between consonants, e.g. መልስ (*mäls*/answer) from ምልስ (*mls*);
- *adjectives* by suffixing various types of bound morphemes, e.g. ደግነት (*dägĭnät*/kindness) from ደግ (*däg*/kind);
- *stems* by prefixing or suffixing various bound morphemes, e.g. ውጤት (*wĭṭet*/result) from ውጥ- (*wĭṭ-*); and
- *nouns* by suffixing various bound morphemes, e.g. ልጅነት (*lĭjĭnät*/childhood) from ልጅ (*lĭj*/child).

Adjectives are also derived from:

- *verbal roots* by infixing vowels between consonants, e.g. ጥቁር (*ṭĭqur*/black) from ጥቅር (*ṭqr*);
- *nouns* by suffixing bound morphemes, e.g. ጥቁር (*ṭĭqur*/black) from ጥቅር (*ṭqr*); and
- *stems* by suffixing bound morphemes, e.g. ደካማ (*däkama*/weak) from ደካም- (*dekam-*).

In addition, nouns and adjectives can be derived from compound words of various lexical categories. Amharic verb inflection is even more complex than that of nouns and adjectives as verbs are marked for any combination of person, gender, number, case, tense/aspect, and mood resulting in the synthesis of thousands of words from a single verbal root. With respect to the derivation process, several verbs in their surface forms are derived from a single verbal stem, and several stems are derived from a single verbal root. For example, from the verbal root ሰብር (*sbr*/to break), we can derive verbal stems such as ሰብር (*säbr*), ሰበር (*säbär*), ሳብር (*sabr*), ሰባብር (*säbabr*), ተሰባብር (*täsäbabr*), etc. and we can derive words such as ሰበረው (*säbäräw*), ሰበርኩ (*säbärku*), ሰበረች (*säbäräč*), ሰበርን (*säbärn*), አሰበረ (*assäbärä*), ተሰበረ (*täsäbärä*), አልሰበረም (*alsäbäräm*), ሲሰበር (*sisäbär*), ሳይሰበር (*saysäbär*), ካልተሰበረ (*kaltäsäbärä*), የሚሰበር (*yämisäbär*), etc. This leads a single word to represent a complete sentence constructed with subject, verb and object. Because of such morphological complexities, many Amharic natural language processing applications require stemmer or morphological analyser as a key component.

3 The Proposed Semantic Network Model

The model proposed to construct Amharic semantic networks has the following major components: *Amharic WordNet*, *text analysis and indexing*, *computing term vectors*, *concept extraction*, and *relation extraction*. First, index terms representing text corpus are extracted. Term vectors are then computed from the index file and stored using WordSpace model. By searching the WordSpace, semantically related concepts are extracted for a given synset in the Amharic WordNet. Finally, relations between those concepts in the intervening word patterns are extracted from the corpus using pairs of concepts from Amharic WordNet. Process relatioships between these components are shown in Figure 1.

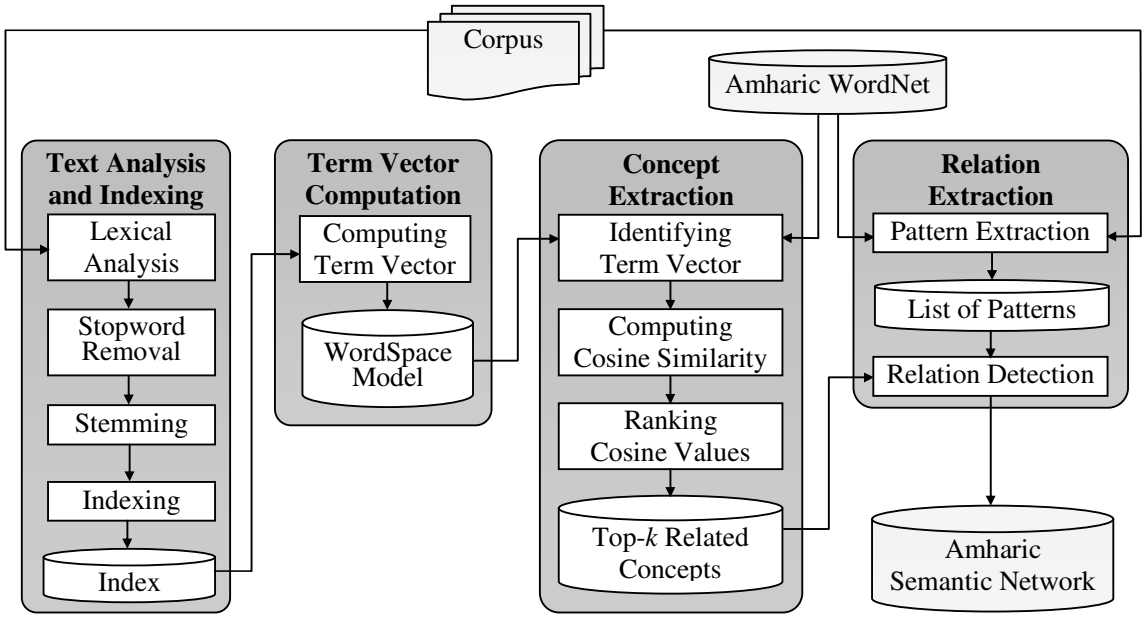

Figure 1. System architecture of the proposed Amharic semantic network.

3.1 Amharic WordNet

To automatically construct semantic networks from free text corpus, we need some initial knowledge for the system so that other unknown relation instances can be extracted. Accordingly, we constructed Amharic WordNet manually as a small knowledge base in which the basic relation between terms is "synonymy". Amharic WordNet is composed of 890 single word terms (all are nouns) grouped into 296 synsets (synonym groups) and these synsets are representations of the concepts of 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", "part-of" and "antonym". The Amharic WordNet is then used to set different seeds for a specific relation. Once we prepare sets of seeds from the WordNet, we can extract the patterns which indicate how these pairs of seeds exist in the corpus. The way these pairs of concepts exist in the corpus can tell us more about other concept pairs in the corpus. 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/Africa} can exist in same way as the former pairs. The patterns extracted between a pair of terms {ኢትዮጵያ/Ethiopia, አፍሪካ/Africa} can be used to extract the relation between other countries like ኬንያ/Kenya with that of አፍሪካ/Africa.

3.2 Text Analysis and Indexing

The process of text analysis starts with removal of non-letter tokens and stopwords from the corpus. This is followed by stemming of words where several words derived from the same morpheme are considered in further steps as the same token. Since Amharic is morphologically complex language, the process of finding the stem which is the last unchangeable morpheme of the word is a difficult task. We used a modified version of the stemmer algorithm developed by Alemayehu and Willet (2002) which removes suffixes and prefixes iteratively by employing minimum stem length and context sensitive rules. The stem is used as a term for indexing which is performed by applying term frequency-inverse document frequency weighting algorithm.

3.3 Computing Term Vectors

A term vector is a sequence of term-weight pairs. The weight of the term in our case is the co-occurrence frequency of the term with other terms in a document. Term vectors are computed from the index file where we extract the co-occurred terms and compute the term vectors in the WordSpace model. 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 using random projection algorithm (Fern and Brodley, 2003). At the end, the WordSpace contains the list of term vectors found from the corpus along with co-occurrence frequencies of each term. The algorithm used to compute term vetcors is shown in Figure 2.

Input:
INDEX=folder containing index files
DIM=dimension of each term vector
Output:
TERMVECTORS=set of term vectors
Initialize:
TERMVECTORS={}

1. Open **INDEX**
2. Create term-document matrix **MATRIX** from **INDEX**
3. For each document **DOC** in **MATRIX**:
   Create basic random vector **RANDVEC**
   For each term **T** in **DOC** and dimension<=**DIM**:
   FREQ= frequency of **T** within **DOC**
   TV= {} //term vector
   TV= **TV** + (**FEQ***RANDVEC)
   Normalize **TV**
   Add **TV** to **TERMVECTORS**
4. Return **TERMVECTORS**

Figure 2. Algorithm for computing term vectors.

3.4 Concept Extraction

Semantically related concepts for a seed term of Amharic WordNet are extracted from the WordSpace model which is used to create a collection of term vectors. 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 similarity between the term vector containing this concept and the remaining term vectors of the WordSpace model. For each term vector TV_i in the WordSpace model and a term vector TVx that corresponds to the synset, the cosine similarity C is computed as:

$$C = \frac{\sum_{i=1}^{n} TVx * TV_i}{\sqrt{\sum_{i=1}^{n} TVx^2 * TV_i^2}} \qquad (1)$$

where n is the number of term vectors in the WordSpace model. Since the collection of the term vectors in the WordSpace is many in number, we rank related terms using the cosine values in decreasing order for selection of top-k number of related concepts for the given synset where k is our threshold used to determine the number of related concepts to be extracted.

3.5 Relation Extraction

The relations among concepts considered in this work are "part-of" and "type-of". We use semi-supervised approach to extract relations where a very small number of seed instances or patterns from Amharic WordNet are used to do bootstrap learning. These seeds are used with a large corpus to extract a new set of patterns, which in turn are used to extract more instances in an iterative fashion. In general, using Amharic WordNet entries, intervening word patterns for a specific relation are extracted from the corpus. For each pair of concepts (C_1, C_2) of which we know the relationship contained in Amharic WordNet, we send the query "C_1" + "C_2" to the corpus. The returned text snapshot is processed to extract all n-grams (where n is set empirically to be $2 \le n \le 7$) that match the pattern "C_1X*C_2", where X can be any combination of up to five space-separated word or punctuation tokens. Thus, "C_1X*C_2" is a pattern extracted from the corpus using concept pair (C_1, C_2) from Amharic WordNet of specific relation. For instance, assume the Amharic WordNet contains the concepts "ኢትዮጵያ (*ityoPya*/Ethiopia)" and "አማራ (*amara*/Amhara)" with "ኢትዮጵያ/Ethiopia" being a hypernym of "አማራ/Amhara". The method would query the corpus with the string "ኢትዮጵያ/Ethiopia" + "አማራ/Amhara". Let us assume that one of the returned text snapshot is "...በኢትዮጵያ ከሚገኙ ክልሎች መካከል አማራ አንዱ ሲሆን... (*...bä'ityoPya kämigäñu kǐlǐloč mäkakäl amara andu sihon...*) ". In this case, the method would extract the pattern "...በኢትዮጵያ ከሚገኙ ክልሎች መካከል አማራ... (*...bä'ityoPya kämigäñu kǐlǐloč mäkakäl amara...*)". This pattern would be added to the list of potential hypernymy patterns list with "ኢትዮጵያ/Ethiopia" and "አማራ/Amhara" substituted with matching placeholders, like "**var1** ከሚገኙ ክልሎች መካከል (*kämigäñu kǐlǐloč mäkakäl*) **var2**". Once the patterns are extracted, the final step is to detect if there is a relation between every pair of concepts extracted from the WordSpace. If a relation between a pair of concepts are detected, the concept pair will be

added to the network in which each concept is a node and the link is the relation between the concepts. Figure 3 shows the algorithm used to extract relations between concepts.

```
Input:
    CONPAIR = set of pair of concepts from
              Amharic WordNet
    CORPUS  = corpus
    PATLIST = list of patterns extracted
Output:
    SEMNET  = list containing pairs of concepts
              with relation R as a form of
              R(C1,C2)
Initialize:
    MODPAT= {}    //list of modified patterns

1. For each pair (A,B) in CONPAIR:
       For each pattern PAT in PATLIST:
          NEWPAT=A + PAT + B
          MODPAT= MODPAT + NEWPAT
2. For each phrase PHRASE in MODPAT:
       COUNT =0
       A= PHRASE [0]
       B= PHRASE [size(PHRASE)-1]
       For each file FILE in CORPUS:
          For each sentence SENTENCE in FILE
             If PHRASE exists in SENTENCE
                COUNT=COUNT+1
             If COUNT>=TRESHOLD
                Add pair (A,B) to SEMNET
                Break
3. Return SEMNET
```

Figure 3. Algorithm for Relation Extraction.

4 Experiment

4.1 Corpus Collection

The corpus is composed of domain independent, unconstrained and unstructured text data. It contains two groups of text. The first group is a collection of news text documents gathered by Walta Information Center (1064 news items) and all news items are tagged with part-of-speech categries.This group of the dataset was used for the extraction of concepts in the corpus. The second group was collected from Ethiopian National News Agency (3261 news items). This dataset group was used for computing the frequency of concepts that are extracted from the first tagged dataset. Thus, a total of 4325 Amharic news documents were collected to build the corpus.

4.2 Implementation

The proposed model was implemented by creating the WordSpace from the index file which is mapped to term-document matrix. We used Apache Lucene and Semantic Vectors APIs for indexing and development of the WordSpace model, respectively. Concept and relation extraction processes were also implemented using Java.

4.3 AMSNet

We coined the name AMSNet to semantic networks automatically constructed using our system from Amharic text. AMSNet consists of a set of concepts and a set of important relationships called "synonym", "part-of" and "type-of". It holds entries as a form of first order predicate calculus in which the predicate is the relation and the arguments are concepts. AMSNet 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. The growing network is not intended to be a complete model of semantic development, but contains specific relations that can be extracted and connected between concepts of the given corpus. Semantic networks not only represent information but also facilitate the retrieval of relevant facts. For instance, all the facts related to the concept "ኢትዮጵያ/Ethiopia" are stored with pointers directed to the node representing "ኢትዮጵያ/ Ethiopia". Another example concerns the inheritance of properties. Given a fact such as "አገር ሁሉ መንግስት አለው (*agär hulu mängĭst aläw*/each country has a government)", the system would automatically conclude that "ኢትዮጵያ መንግስት አላት (*ityoPya mängĭst alat*/Ethiopia has a government)" given that ኢትዮጵያ አገር ናት (*ityoPya agär nat*/Ethiopia is a country).

4.4 Test Results

There is no gold standard to evaluate the result of semantic network construction. Our result is validated manually by linguists, and based on their evaluations the average 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. Sample result generated from the our system is shown in Figure 4.

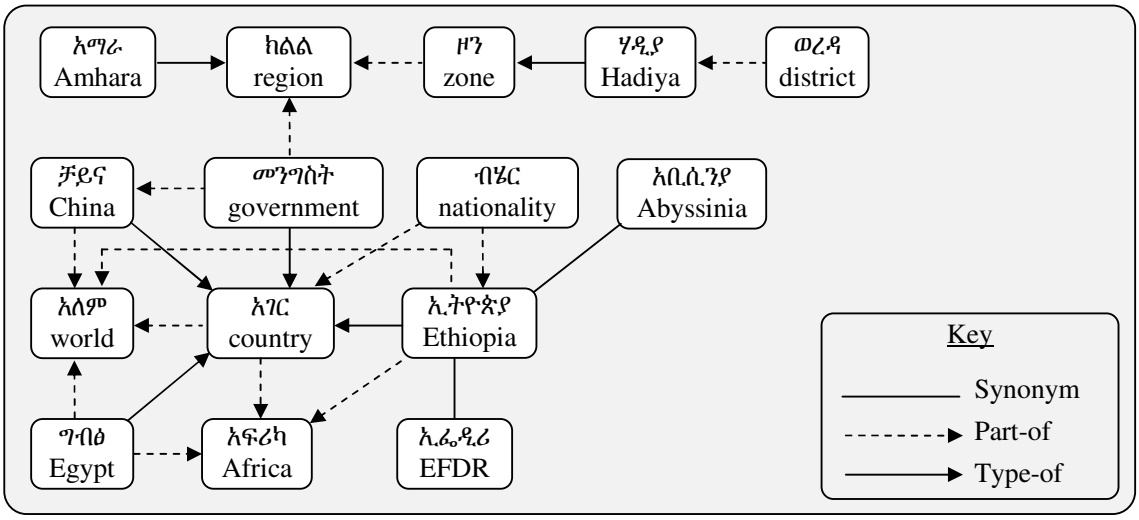

Figure 4. Part of the Amharic semantic network automatically constructed by the proposed system.

5 Conclusion and Future Works

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 availability of Amharic semantic networks helps other Amharic NLP applications such as information retrieval, document classification, machine translation, etc. improve their performance. Future works include deep morphological analyis on Amharic and the use of hybrid approches to improve the performance of the system.

References

Nega Alemayehu and Peter Willet. 2002. Stemming of Amharic Words for Information Retrieval, In *Literary and Linguistic Computing*, Vol 17, Issue 1, pp. 1-17.

Getahun Amare. 2010. ዘመናዊ የአማርኛ ሰዋስው በቀላል አቀራረብ (Modern Amharic Grammar in a Simple Approach). Addis Ababa, Ethiopia.

Tim Berners-Lee. 2001. The Semantic Web, *Scientific American* , Vol 284, Issue 5, pp. 34-43.

Widdows Dominic and Cohen Trevor. 2010. The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics, In *Fourth IEEE International Conference on Semantic Computing (IEEE ICSC2010)*. Carnegie Mellon University, Pittsburgh, PA, USA.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MIT Press.

Tsatsaronis George, Iraklis Varlamis and Michalis Vazirgiannis. 2010. Text Relatedness Based on a Word Thesaurus, *Journal of Artificial Intelligence Research*, vol. 37, pp. 1-39.

Brian Harrington and Stephen Clark. 2007. ASKNet: Automated Semantic Knowledge Network, In *Proc. 22nd National Conf. on Artificial Intelligence*, Vancouver, Canada. pp. 889-884.

Paul Lewis, Gary Simons and Charles Fennig 2013; *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International.

Stephen Richardson, William Dolan and Lucy Vanderwende. 1998. MindNet: Acquiring and structuring semantic information from text, In *Proceedings of the 17th COLING*, Montreal, Canada. pp. 1098-1102.

Magnus Sahlgren. 2006. The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector space. *PhD Thesis*, Stockholm University, Sweden.

Andrew Smith. 2003. Automatic Extraction of Semantic Networks from Text using Leximancer, In *Proceedings of HLT-NAACL*, Edmonton.

Mark Steyvers and Joshua Tenenbaum. 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth, *Cognitive Science,* Vol 29, Issue 1, pp. 41–78.

Xiaoli Fern and Carla Brodley. 2003. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach, In *Proc. of the 20th Int. Conf. on Machine Learning (ICML-2003)*, Washington, DC.

Baye Yimam. 2000. የአማርኛ ሰዋስው (Amharic Grammar). Addis Ababa, Ethiopia.