The Need for Amharic WordNet

Tessema Mindaye
Computer Science Department
Addis Ababa University
tessemin@cs.aau.edu.et

Meron Sahlemariam
IS&T Division
UN ECA
nahmmer@gmail.com

Teshome Kassie
Ministry Finance and Economic
Development, Ethiopia
tkheran@yahoo.com

Abstract

WordNet has been recognized as a valuable resource in the human language technology and knowledge processing communities. Due to the success of Princeton WordNet, many language specific WordNets have been developed and are still in development. In this paper the need for Amharic WordNet is discussed and a way forward is also suggested.

1 Introduction

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called *synsets*, provides short, general definitions, and records the various semantic relations between these synsets. WordNet has become one of the most valuable resources for a wide range of Natural Language Processing (NLP) research and applications, such as automatic word-sense disambiguation, information retrieval and document summarization and clustering. Due to the success of Princeton WordNet (PWN), many language specific WordNets have been developed and are still in development. Despite its application, there is no Amharic WordNet so far.

This paper is organized as follows; section 2 discusses the Amharic language, the script it uses and typical features of the language. Section 3 discusses different Amharic tools and the applications of Amharic WordNet for those tools. Section 4 discuses future work and gives conclusions.

2 The Amharic Language

Ethiopia is a linguistically diverse country where more than 80 languages are used in day-to-day communication. Although many languages are spoken in Ethiopia, Amharic is dominant in that it is spoken as a mother tongue by a substantial segment of the population and it is the most commonly learned second language throughout the country (Marvin et al., 1976). The language is the official language of the federal government of the country. According to the 1998 census of the country (ECSA, 1998), Amharic is the first language of more than 17 million people and second language for more than 5 million people.

2.1 The Amharic Writing System

According to Marvin et al. (1976), three writing systems are in use in Ethiopia, the Amharic syllabary, the Roman alphabet, and Arabic script. The Amharic syllabry, which is derived from the writing system of ancient South Arabian inscriptions, is used for Ge'ez, Amharic, and Tigrigna, with slight modification. The Amharic syllabry is uniquely Ethiopian writing system. 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 attested in inscriptions in the Sabean dialect (Marvin et al., 1976). The original Sabaean alphabet is said to have had 29 symbols. When Ge'ez became the spoken and written language in common use in northern Ethiopia, it took only 24 of the 29 Sabaean symbols, modify most of them and add two new symbols to represent sounds of Greek and Latin loanwords not found in Ge'ez. The style of the writing was also modified to left to right. By the time Ge'ez ceased to be a living spoken and written language and replaced by Amharic and other languages, further changes took place. Amharic did not discriminate in adopting the Ge'ez fidel; it took all of the symbols (Yemam, 1987) and added some new ones that represent sounds not found in Ge'ez. These added alphabetic characters are ቸ, ጨ, ጀ, ኝ, ሸ, ኸ, and ፐ.

Currently, the language's writing system contains 34 base characters each of which occurs in a basic form and six other forms known as orders. The seven orders represent syllable combinations consisting of a consonant following vowel. This is why the Amharic

writing system is often called syllabic rather than alphabetic, even if there is some opposition (Yemam, 1987). The 34 basic characters and their orders give 238 distinct symbols. In addition, there are forty others that contain a special feature usually representing labialization e.g.ጒ, ቈ. In Amharic there is no Capital-Lower case distinction. There are also punctuation marks and numeration system.

2.2 Typical characteristics of Amharic Language

There is a process of change in any language in many of its aspects: change of meaning, change of syntax, phonetic change, etc (Haile, 1967). The case for Amharic is not different; especially the script underwent changes when it was borrowed from Ge'ez. Through the adaptation process and other factors the Amharic writing system got some problems.

The first problem is the presence of "unnecessary" alphabets (fidels) in the language's writing system. These fidels (alphabets) have the same pronunciation but different symbols. These different fidels can be used interchangeably without meaning change. The fidels are ሀ and ዐ, ኀ and ሀ, ሰ and ሠ and ሀ, ሐ, and ኸ. For example, the word "sun" can be written as, ጸሀይ, ጸሃይ, ዐሃይ, ዐሃይ, etc … all mean the same, although they are written differently.

The other problem is in the formation of compound words. 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". There are many such compound words, which need some effort to have a standard way of forming them.

Amharic is morphologically rich language where up to 120 words can be conflated to a single stem (Alemayehu and Willet, 2002). The word units of Amharic are phoneme, morpheme, root, stem, and word. The 34 base characters are a phoneme. A collection of phonemes forms morphemes, which is the smallest meaningful unit in a word (Yemam, 1987). 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.

In Amharic language, it is common to write some words in shorter form using "/" (forward slash) or "." (dot). The short form of words can

be expanded as single or a combination of words. አ/አ, which is expanded as አዲስ አበባ (means Addis Ababa), is an example for the latter. መ/ር is a short form of the single word መምህር (means teacher).

Another problem of the language is, 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; "ሂጃ" vs. "ሂጅ", "አይዶላም" vs. "አይደላም", "ጓጤ" vs. "ጓጊ" , etc (Yacob, 2006). 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, ኮምፒዉተር, etc.

3 Application of Amharic WordNet

Tools that are developed for Amharic language need to consider the above-mentioned characteristics of the language in one way or another. Some of these tools are discussed below together with how the use of Amharic WordNet can increase their performance. The tools are developed without the use of Amharic WordNet.

3.1 Amharic Search Engine

The Web is a huge repository of information in the form text, image, audio, and video. Search engines, such as Google, Yahoo!, etc, are the first port of call for the discovery of resources from this huge repository. According to Internet World Stats, usage and Population (2009) Ethiopia took 0.4 % of Internet users out of Africa's share in 2009. The statistics also shows that there was an increase of users of Internet in Ethiopia by 3500% during the years 2000-2009. Due to this increase in Internet population within the country and large number of population that speaks the language in Diaspora, the number of web documents that are written in Amharic language and Ethiopic script is increasing. In order to search these documents we need a search engine that can handle Amharic queries, written in Ethiopic script, well.

As described earlier, Amharic has many unique features that affect the retrieval of the language's documents from the Web. ሐበሻ search engine (Mindaye, 2007) is a complete

language specific search engine that is developed for Amharic web documents. The search engine has three components: the Amharic Crawler, The Amharic Indexer and the Amharic Query Engine. The crawler (Language Specific Crawler) crawls the Web and collects Amharic web documents and stores them in a repository. The Indexer processes the documents and stores them in a structure that is efficient for searching. Some of the processing in this component are: tokenization, stop word removal, stemming, etc. The Query Engine component gives an interface that the user can enter his/her information need in Amharic language using Ethiopic script. It returns the relevant documents according to their rank.

The application of Amharic WordNet is many folds in ሐበሻ search engine. Queries can be expanded using Amharic WordNet that will increase the recall of the search engine. Interchangeable alphabets (repetitive alphabets) can also handled easily by Amharic WordNet by considering all the different forms of a word as a synonym. The same applies for the regional dialects, Loan words and short form of a compound word, which are all different ways of writing a same word as discussed in section 2.

3.2 Amharic Automatic Text Categorization

The process of automatic text categorization involves calculating similarities between documents and categories using the information extracted from the document. In recent years, ontology-based document categorization method is introduced to solve the problem of document classifier. In order to resolve the problem of not considering semantic relationships between words, one study (Sahelemariam et al., 2009) proposes a framework that automatically categorizes Amharic documents into predefined categories using knowledge represented in the News ontology. At the heart of its classification system is the knowledge base that enables the representation of different domain concepts. With the help of News domain ontology, this study categorizes a given Amharic document (news) into a specific predefined category. The study shows that the use of concepts for Amharic document categorizer obtained a promising outcome. However, the study also recommend for further research and

developmental effort in the area of external knowledge base such as Amharic WordNet. In the study, for the process of extracting concepts from the knowledge base, index terms are mapped on the corresponding concepts of the ontology. However, there is a possibility that the term may not exist because of the limited number of concepts available in the News ontology. This situation requires an alternative way of mapping onto the external knowledge base concept. The alternative way is to use the extended concept in order to map between the external concept and the existing knowledge base. In the linguistic knowledge base, we can find words semantically related with the other words in many ways. Using various semantics relationships, we can take the advantage to establish links between the words of the ontology concepts and WordNet vocabulary. These retrieved words from WordNet will be treated as the external knowledge base to enhance the result of the study.

3.3 Amharic Word Sense Disambiguation

Ambiguity is defined as the property of being ambiguous, where a word, term, notation, sign, symbol, phrase, sentence, or any other form used for communication, is called ambiguous if it can be interpreted in more than one way (Mihalcea and Pedersen, 2005). When language is capable of being understood in more than one way by a reasonable person, ambiguity exists. Ambiguity is inherent to human language. Successful solutions for automatic resolution of ambiguity in natural language often require large amounts of annotated data/knowledge resources to achieve good levels of accuracy. One study tries to develop a tool for Amharic word sense disambiguation (Kassie, 2009). In the study, Amharic Penal Code document was used for experimentation by applying Semantic Vectors of words of dimension 200. The term vectors are built from index of terms using lucene IR library. Using those term vectors thesaurus can be constructed by calculating the k nearest neighborhood from the word space by applying the distance measure between points of term representation according to the usage of terms in documents. In other words, a query that is one word is run using the prototype where the system retrieves words by applying the similarity calculation of nearest neighborhoods from documents according to their usage. The neighborhood is

calculated from the co-occurrence frequency of words in documents. The average precision and recall of the system is 58% and 82% respectively. The author argued that if there was an Amharic WordNet the performance of the tool definitely would improve.

4   Conclusion and Future Work

WordNet has been recognized as a valuable resource in the human language technology and knowledge processing communities. From the above sections we can clearly see the application of Amharic WordNet. Developing the WordNet will enhance the performance of many information retrieval and natural language processing tools for the language. It will also give the language a chance to be integrated with other languages for cross-language processing.

Princeton WordNet is a great inspiration for the development of WordNet in different languages. There have been   many efforts o develop a WordNet for different languages such as Arabic (Elkateb et al., 2006), different European languages (Vossen, 1997), etc. There are two approaches of developing a WordNet: the *merge* approach and *extended* approach. In order to develop Amharic WordNet, an extended approach seems appropriate due to the following reasons:

- It reduces the cost and time of developing Amharic WordNet from scratch.
- It gives an opportunity to integrate the language WordNet with other languages WordNet.
- It is wise to use the information in the Princeton WordNet for such under-resourced languages like Amharic.

However we may need to modify the PWN in order to incorporate some unique futures of Amharic language. This indicates the need for a coordinated effort from a linguist, Computer Science professionals and other stakeholders to develop Amharic WordNet (AmWN).

References

Baye Yemam. 1987 ዓ.ም. የአማርኛ፣ ሰዋስው። ት.መ.ማ.ማ.ድ. .።

Daniel Yacob. 2006. "*Application of the Double Metaphone Algorithm to Amharic Orthography*", International Conference of Ethiopian Studies.

Elkateb, S., Back, W., Vossen, P., Farwell, D., Rodrigue, H., Pease, A., Alkhalifa, M. and Fellbaum, C. 2006. *Arabic WordNet and the Challenges of Arabic. The Challenges of Arabic for NLP/MT*. International Conference at the British Computer Society (BSC), London.,

Ethiopian Central Statistical Authority (ECSA). 1998. *The 1994 Population and Housing Census of Ethiopia: Results at Country Level.* Vol.1, Statistical Report 44, AddisAbaba, Ethiopia.

Getachew Haile. 1967. *The Problems of AmharicWriting System*. Unpublished.

Internet World Stats, Usage and Population Statistics. 2009. Available at: http://www.InternetworldStats.com/stats.htm

Marvin L. Bender, Head W. Sydeny, andRoger Cowley. 1976. *The Ethiopian Writing System*. In Bender et al (Eds.) Language in Ethiopia. London, Oxford University press.

Meron Sahlemariam, Mulugeta Libsie, and Yacob, Daniel. 2009. "*Concept-Based Automatic Amharic Document Categorization". AMCIS 2009 Proceedings.* Paper 116. http://aisel.aisnet.org/amcis2009/116

Nega Alemaehu and Willet P. 2002. *Stemming of Amharic Words for Information Retrieval*. In Literary and Linguistic Computing. Oxford, Oxford University press, Vol. 17, No.1, pp 1-17.

Rada Mihalcea and Ted Pedersen. 2005. .*Advances in Word Sense Disambiguation Tutorial* at AAAI-

Teshome Kassie. 2009. *Word Sense disambiguation for Amharic Text Retrieval: A Case Study for Legal Documents*. Thesis ,Addis Ababa University.

Tessema Mindaye. 2007. *Design and implementation of Amharic Search Engine* .Masters Thesis ,Addis Ababa University.

Vossen P. 1997. *EuroWordNet: a multilingual database for information retrieval*. In: Proc. of the DELOS workshop on Cross-language Information Retrieval, Zurich, Switzerland.