Machine Translation for Amharic: Where we are

Saba Amsalu*[*] & Sisay Fissaha Adafre[†]

[*]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld
Kiskerstrasse 6, Bielefeld
Tel: 0049 (0)521 1063519
{saba@uni-bielefeld.de}
[†]Informatics Institute, University of Amsterdam
Tel: 0031205257545
{sfissaha@science.uva.nl}

Abstract

We describe some of the efforts in research and resource acquisition towards developing an Amharic machine translation system. A brief account of the challenges of unavailability of text corpora and linguistic tools is presented. In the direction of research, our survey shows that only limited attempts to integrate Amharic into a unification based machine translation system and to extract lexical information from bilingual corpora which are the basis for machine translation have been made. In order to fill the gap, significant research on different aspect of the language needs to be done. In this paper we identify and describe some of the tasks that need to be given due attention both in terms of resource acquisition and developing tools. This way we hope to inspire ideas and more discussions on the issue.

1. Introduction

Machine translation (MT) systems have shown to provide significant economic gains in a number of areas. Now that notable progress has been made in the development of MT systems for economically important languages of the world, the focus has shifted recently to the design of methods for the rapid development of MT system for minority languages which have no or very little natural language processing resources. Amharic is one of such languages for which limited research has been carried out in the area of MT or NLP in general.

In this paper, we provide a short summary of some of the works done in the areas of Amharic natural language processing, which are essential to the development of machine translation system for Amharic. Sections 2. provides background information on Amharic language. Section 3. briefs the need and challenges for developing MT system and the resources and tools available. Sections 4. & 5. describe relevant research towards having an MT system and future directions in corpus acquisition developing relevant tools.

2. Amharic Language

Amharic is a Semitic language which is widely spoken in Ethiopia. It has 17.4 million native and around 4 million non-native speakers and a long history of service as a medium in education and government activities (Wikipedia, 2006). It has its own script (Fidel) that is borrowed from Geez, another Ethiopian Semitic language (Leslau, 1995). The script is a syllabary writing system where each character represents an open CV syllable.

Amharic has a complex morphology. Word formation involves prefixation, suffixation, infixation, reduplication, and Semitic stem interdigitation, among others. Like other Semitic languages, Amharic verbs and their derivations constitute a significant part of the lexicon. In Semitic languages, words, especially verbs, are best viewed as consisting of discontinuous morphemes that are combined in a non-concatenative manner. Syntactically, Amharic is an SOV language.

3. Available Data and Tools to support MT

The need for automatic means of processing Amharic language has recently been recognised (Bayou, 2000; Bayu, 2002; Alemayehu and Willett, 2002; Getachew, 2001; Fissaha, 2004a; Adafre, 2004b; Amsalu and Gibbon, 2005; Alemu, 2005). Most of these studies are limited in scope and are far from meeting the demands of modern NLP applications such as MT. Unavailability of text in machine readable form, the limited number of researchers working in the area and the economic limitations to support projects in resource acquisition and developing tools are some of the major problems facing the development of Amharic MT. Despite these short-comings, the recent efforts coupled with the increasing availability of information technology for Amharic resulted in useful resources.

Amharic Script

Advances have been made with respect to the representation and manipultation of Amharic script using computers. Several fonts and encoding schemes have been proposed for the script in the past, which resulted in the proliferation of non-standard software packages. Currently, there are unicode compatible fonts, and companies are updating their packages to this standard.

The creation of the Unicode standard for Ethiopic was an important step for computational linguists in that certain problems of incompatibility to operating systems and the need to transliterate text from different packages to a common representation in Latin script has been alleviated.

Text Corpora

Historically, Amharic has served as the national language of Ethiopia. Currently it is being used as the working language of the Federal Government and several local governments. This has resulted in a large quantity of publications which range from news articles, working documents in organisations and novels to religious manuscripts and teaching materials for elementary schools. Among these documents, news articles, the Bible, government documents such as constitutional papers, judiciary documents, forms of Banks and Insurance and other companies are available in bilingual versions; and in most cases in machine readable form. The rest are mostly monolingual print documents.

Bilingual Dictionaries

There are different print bilingual dictionaries (Amharic-English). Some of these dictionaries are also available in machine readable form (Aklilu, 1998; Leslau, 1996; Davidovic Mladen, 1992). There are also online dictionaries (Eliab, ; Cain,). Though these dictionaries may have limited coverage both in the lexical entries they contain and the lexical data categories for the entries, they can be useful for bootstrapping large scale dictionaries or other linguistic resources such as wordlists and thesauri.

Linguistic Tools

Studies in the languages of Ethiopia, particularly Amharic, have been mainly motivated by pure linguistic interest (Baye, 1986; Leslau, 1995; Bender and Hailu, 1978). Most of the works are of descriptive nature and focus mainly on the morphology of the language. Some of the rare attempts at formalising Amharic grammar have been carried out within the framework of early theories of transformational grammar and are limited in scope. Regarding the computational aspect, it is only recently that works have begun to apply some of the techniques of formal linguistics (Bayou, 2000; Bayu, 2002; Alemayehu and Willett, 2002; Getachew, 2001; Fissaha, 2004a; Adafre, 2004b; Amsalu and Gibbon, 2005). The results of most of these works are prototypes with limited scope. To the best of our knowledge, there are no wide coverage parts-of-speech taggers, morphological analysers or syntactic parsers for Amharic. Recognising these problems, some recommendations have been made to develop resources required for the development of these tools such as Amharic treebank (Alemu et al., 2003).

4. Related Research

With respect to MT, one of the early attempt to integrate Amharic into a unification based machine translation system is work done by (Fissaha and Haller, 2003a). This work provided formal description of Amharic language borrowing ideas from contemporary linguistic theories, and applied different natural language tools and techniques for solving some of the problems of Amharic languages. Primarily, Xerox finite state tools (XFST) are used for modelling Amharic morphology. Amharic syntax is described using a unification-based grammar formalism. A fragment of the grammar thus developed have been used to develop a prototype transfer-based Amharic-English machine translation system. Corpus-based methods such as collocation extraction, clustering and classification techniques were applied for lexical development (Adafre and Haller, 2003b). However, as with other related researches, this work was severely limited by the inadequacy of monolingual linguistic researches and resources available for Amharic at the time.

Recently, there is increasing interest in the extraction of bilingual lexicon from parallel corpora which we briefly summarise below.

4.1. Bilingual Lexicon Extraction

Development of bilingual lexicon constitutes an important and achieveable short term goal that contributes greatly to machine translation research activities for Amharic. Bilingual lexicon acquisition from Amharic-English parallel corpora, using the Bible as a data source have been given due attention recently (Atelach Alemu and Eriksson, 2004; Amsalu, 2006a; Amsalu and Gibbon, 2006; Amsalu, 2006b).

Atelach Alemu and Eriksson (2004) devised a method for identifying noun translation equivalents from Amharic-English bilingual corpus. They used statistical method with and without the use of an affix stripper for Amharic.

Several modules that work independently and claim to have reasonable degree of success have also been developed by Amsalu and Gibbon (2006):

1. Analysis of term distribution in text for 1:1 alignment
2. Analysis of context of terms for m:n alignment
3. Use of relatively fixed realization of keywords as anchor for alignment
4. Use of syntactic location for parsing Amharic verbs

These modules use the distributional properties of lexical items in parallel corpora and characteristics of syntactically fixed expressions. Promising results have been achieved by modules 1, 2 and 3 (Amsalu, 2006a; Amsalu, 2006b). Some preliminary results have also been obtained by module 4.

5. Future Work

We believe that significant work needs to be done at all levels of Amharic NLP in order to bring about meaningful change to the current status. Essentially, due attention needs to be given to aspects that we broadly categorise as corpus acquisition and developing linguistic tools. Subsequently, we forward tractable approaches which we also believe are applicable for languages in similar situation.

5.1. Corpus Acquisition

Large text corpora form the basis of many monolingual and multilingual research in natural language processing, ranging from developing multilingual lexicons to statistical machine translation systems. Apparently, collecting text corpora written in different languages constitutes an important prerequisite for these research activities. Some of the tasks that can be done in this respect are:

1. Exploit the web: Automatic or semi-automatic acquisition of available corpora from the Web is an easy way to obtain free data. Mainly, newspaper archives are available in large quantities.
2. Collaborative content development (Wiki): Amharic is one of the languages for which a free encyclopedia, i.e. Wikipedia, is being created by Wikimedia foundation (Wikipedia, 2006). Though the current content of Amharic Wikipedia is very small, it has a potential of enabling a rapid development of Amharic corpus; provided that adequate awareness is created among

Amharic speaking community. Any native Amharic speaker with a working knowledge of English can translate English Wikipedia pages into Amharic versions. Wikis provide a general framework in which people on the Web can collaboratively develop content. Recognising this fact, Yacob (2006) has created an Amharic Wiki site where people can share ideas and contribute resources.

3. Exploiting data from other sources: An integrated approach of using OCR system for scanning print documents (there are some attempts to develop OCR systems (Cowell and Hussain, 2003; Alemu, 2005)), gathering print versions of anything available from authors or publishers, and organising projects for manually encoding documents into electronic format.

4. Developing programs for data conversion: To have data of longer period of time, tools that convert text written in non-standard packages need to be developed. This surely is a necessary task and less expensive.

5.2. Developing Tools

A strategic approach of developing tools that enable fast production of machine translation systems is of utmost importance. We propose some useful strategies as follows:

1. Adopting tools developed for related languages: Amharic shares a number of common linguistic properties with Arabic and Hebrew for which active research is being carried out. Use of resources developed for these languages may speed up some of the efforts on Amharic NLP.

2. Machine learning approaches to language modelling: Exploring application of unsupervised machine learning methods for Amharic needs to be given due attention.

3. Using one language as a pivot: Focusing into translating text in one language, namely English, and to perform the translation from other languages through this language.

4. Building domain specific translation machines: A phase by phase approach of addressing a sublangauge would be much easier instead of trying to create a general purpose MT system at once.

5. Fast production of unidirectional MT system: Not much is there to translate from Amharic to English, but the reverse is a lot, so developing a unidirectional MT system that translates from English to Amharic would be practical.

6. References

Sisay Fissaha Adafre and Johann Haller. 2003b. Application of corpus-based techniques to amharic texts. In *MT Summit IX Workshop Machine Translation for Semitic Languages: Issues and Approaches*.

Sisay Fissaha Adafre. 2004b. Adding amharic to a unification-based machine translation system.

Amsalu Aklilu. 1998. *English-Amharic Dictionary*. Oxford University Press.

Nega Alemayehu and Peter Willett. 2002. Stemming of amharic words for information retrieva. *Literary and Linguistic computing*, 17(1):1–17.

Atelach Alemu, Lars Asker, and Gunnar Eriksson. 2003. An empirical approach to building an amharic treebank. In *Proceedings of 2nd Workshop on Treebanks and Linguistic Theories*, Vaxjo University, Sweden.

Worku Alemu. 2005. *Handwritten Amharic Character Recognition Applied to Bank Checks*. Ph.D. thesis, Dresden University of Technology.

Saba Amsalu and Dafydd Gibbon. 2005. Finite state morphology of amharic. In *Proceedings of the International Conference on Recent Advances n Natuaral language processing*, pages 47–51, Borovets, Bulgaria.

Saba Amsalu and Dafydd Gibbon. 2006. Methods of bilingual lexicon extraction from amharic-english parallel corpora. In *Proceedings of The 5th World Congress of African Linguistics*, Addis Ababa. to appear.

Saba Amsalu. 2006a. Data-driven amharic-english bilingual lexicon acquisition. In *Proceedings LREC2006*, Genoa, Italy.

Saba Amsalu. 2006b. Scaling up from word to phrasal alignments of amharic-english parallel corpora. Submitted.

Lars Asker Atelach Alemu and Gunnar Eriksson. 2004. Building an amharic lexicon from parallel texts. In *Proceedings of: First Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon and Corpus Compilation, a workshop at LREC*, Lisbon.

Abiyot Bayou. 2000. Developing automatic word parser for amharic verbs and their derivation. Master's thesis, Addis Ababa University, Addis Ababa.

Tesfaye Bayu. 2002. Automatic morphological analyzer for amharic: An experiment involving unsupervised learning and autosegmental analysis approaches. Master's thesis, Addis Ababa University, Addis Ababa.

Matthew Cain. Online dictionary of the official language of ethiopia. `http://www.amharicdictionary.com/`.

John Cowell and Fiaz Hussain. 2003. Amharic character recognition using a fast signature based algorithm. In *Proceedings of Seventh International Conference on Information Visualization (IV'03)*, page 384, Vaxjo University, Sweden.

A. Zekaria Davidovic Mladen. 1992. *Amharic-English / English-Amharic Dictionary*. Hippocrene Books Inc.

Eliab. Online amharic-english dictionary. `http://www.ethiopiandictionary.com/`.

Sisay Fissaha and Johann Haller. 2003a. Amharic verb lexicon in the context of machine translation. *TALN*.

Sisay Fissaha. 2004a. Formal analysis of some aspect of amharic noun phrases. In *EAMT 2004 Workshop*, Malta.

Mesfin Getachew. 2001. Automatic part of speech tagging for amahric language: An experiment using stochastic hidden markov model (hmm) approach. Master's thesis, School of Graduate Studies of Addis Ababa University.

Wolf Leslau. 1996. *Concise Amharic Dictionary*. University of California Press.

Wikipedia. 2006. Amharic language. http://en.wikipedia.org/wiki/Amharic_language.

Daniel Yacob. 2006. Welcome to amharic nlp. http://nlp.amharic.org/.