

Applying Machine Learning to Amharic Text Classification

Björn Gambäck

Magnus Sahlgren

Atelach Alemu Argaw

Lars Asker

Userware Laboratory
Swedish Institute of Computer Science AB
Box 1263, SE-164 29 Kista, Sweden
{gamback,mange}@sics.se

Department of Computer and System Sciences
Stockholm University
Forum 100, SE-164 40 Kista, Sweden
{atelach,asker}@dsv.su.se

Even though the last years have seen an increasing trend in investigating applying language processing methods to other languages than English, most of the work is still done on very few and mainly European and East-Asian languages. However, there is a need for people all over the World to be able to use their own language when using computers or accessing information on the Internet. This requires the existence of a variety of applications including local language spell-checkers, word processors, machine translation systems, search engines, etc. At the same time, the amount of work required to develop all aspects of natural language processing for a new language is huge.

The main obstacles to progress in language processing for new languages are three-fold. Firstly, the peculiarities of a language itself might force new strategies to be developed. Secondly, the lack of already available resources and tools creates a vicious circle: having resources makes producing resources easier, but not having resources makes the creation and testing of new ones more difficult and time-consuming. Thirdly, there is commonly a lack of interest (and understanding) of the needs for people to be able to use their own language in computer applications — a lack of interest both in the surrounding world, but also sometimes even in the countries where a language is used. Until a critical mass of language processing resources have been made publicly available, it is not likely that natural language processing for any language can “take off” and reach a level higher than that of academic prototype systems. Making such resources publicly available will further boost activities in the area by allowing researchers to benefit from the work of others and thereby to build more advanced tools without having to invent the wheel over and over again. We believe this to be the best chance for any language with limited computational linguistic resources to break the vicious circle of always having to start from scratch with research in the area.

Amharic is the language for country-wide communication in Ethiopia and has its own writing system which was incorporated into Unicode only in year 2000. It is quite dialectally diversified and probably representative of the languages of a continent which so far has received little attention within the language processing field. Amharic is a language for which very few computational linguistic tools or corpora (such as lexica, part-of-speech taggers, parsers or tree-banks) exist.¹ This problem, which is shared by a number of so called “under-resourced languages” has proven to be a bottleneck when it comes to promoting the use of computers and the Internet in the language. It is difficult to develop new linguistic resources without access to already existing ones. We are currently involved in project the primary goal of which is to investigate how well existing linguistic knowledge can be transferred between languages and to develop tools and techniques that can support such knowledge transfer. Another goal is to support the collection and publication of a large (mono- and bilingual) corpus and other resources for Amharic.

¹See [1] for an overview of the efforts that have been made so far to develop language processing tools for Amharic. We should note though that some of the previous work most relevant to the present paper were hampered by the lack of annotated data.

The project involves the following main steps:

1. The collection a large corpus of Amharic text with the aim to create a standard dataset for the application of automated methods (e.g., POS tagging), Determine a suitable or appropriate set of tags for the language and manually tag words in the corpus with appropriate lexical attributes. This requires considerable linguistic expertise provided by the Department of Linguistics at Addis Ababa University.
2. Build efficient and accurate part-of-speech taggers and morphological analyser semi-automatically based on the corpus data and available processing tools for other languages.
3. Refine and extend the tagged corpus automatically using the tagger and the morphological analyser. Using a parallel corpus (e.g., for English or French), information can be transferred from resource-rich languages to assist the refinement by supplying disambiguation clues.
4. Manually check and correct the corpus followed by retraining and automatic refinement of the tagger and the morphological analyser given the extended corpus.
5. Create a set of queries pertaining to the document collection and identify automatically (using machine-learning methods) and semi-automatically (with the support of domain experts) which documents are relevant to a given query and which belong to a specific category of text.

The present presentation will describe the entire project and the possibilities it will create for Amharic language processing; however, it will concentrate on the last step, and in particular on how we are using machine learning methods to categorize Amharic text. The corpus of step 1 is currently under development, so we have undertaken some initial experiments on a smaller corpus of Amharic news text. In previous work [2], we discussed doing this by utilizing Self-Organizing Maps (an Artificial Neural Network strategy). Here we will show how those results can be improved using the Random Indexing vector space methodology [3].

1. References

- [1] A. Alemu, L. Asker, and M. Getachew, “Natural language processing for Amharic: Overview and suggestions for a way forward,” in *10th Conf. Traitement Automatique des Langues Naturelles*, Batz-sur-Mer, France, 2003.
- [2] S. Eyassu and B. Gambäck, “Classifying Amharic news text using Self-Organizing Maps,” *43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, Michigan: 2005, workshop on Computational Approaches to Semitic Languages.
- [3] M. Sahlgren and J. Karlgrén, “Automatic bilingual lexicon acquisition using random indexing of parallel corpora,” *Natural Language Engineering*, vol. 11, no. 3, 2005.