Classifying Amharic News Text Using Self-Organizing Maps

Samuel Eyassu
Department of Information Science
Addis Ababa University, Ethiopia
samueleya@yahoo.com

Björn Gambäck[*]
Swedish Institute of Computer Science
Box 1263, SE–164 29 Kista, Sweden
gamback@sics.se

Abstract

The paper addresses using artificial neural networks for classification of Amharic news items. Amharic is the language for countrywide communication in Ethiopia and has its own writing system containing extensive systematic redundancy. It is quite dialectally diversified and probably representative of the languages of a continent that so far has received little attention within the language processing field.

The experiments investigated document clustering around user queries using Self-Organizing Maps, an unsupervised learning neural network strategy. The best ANN model showed a precision of 60.0% when trying to cluster unseen data, and a 69.5% precision when trying to classify it.

1 Introduction

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; for the vast number of languages of the African continent there still remains plenty of work to be done. The main obstacles to progress in language processing for these are two-fold. Firstly, the peculiarities of the languages themselves might force new strategies to be developed. Secondly, the lack of already available resources and tools makes the creation and testing of new ones more difficult and time-consuming.

Many of the languages of Africa have few speakers, and some lack a standardised written form, both creating problems for building language processing systems and reducing the need for such systems. However, this is not true for the major African languages and as example of one of those this paper takes Amharic, the Semitic language used for countrywide communication in Ethiopia. With more than 20 million speakers, Amharic is today probably one of the five largest on the continent (albeit difficult to determine, given the dramatic population size changes in many African countries in recent years).

The Ethiopian culture is ancient, and so are the written languages of the area, with Amharic using its own script. Several computer fonts for the script have been developed, but for many years it had no standardised computer representation[1] which was a deterrent to electronic publication. An exponentially increasing amount of digital information is now being produced in Ethiopia, but no deep-rooted culture of information exchange and dissemination has been established. Different factors are attributed to this, including lack of digital library facilities and central resource sites, inadequate resources for electronic publication of journals and books, and poor documentation and archive collections. The difficulties to access information have led to low expectations and under-utilization of existing information resources, even though the need for accurate and fast information access is acknowledged as a major factor affecting the success and quality of research and development, trade and industry (Furzey, 1996).

[*] Author for correspondence.

[1] An international standard for Amharic was agreed on only in year 1998, following Amendment 10 to ISO–10646–1. The standard was finally incorporated into Unicode in year 2000: www.unicode.org/charts/PDF/U1200.pdf

Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages, pages 71–78,
Ann Arbor, June 2005.

In recent years this has lead to an increasing awareness that Amharic language processing resources and digital information access and storage facilities must be created. To this end, some work has now been carried out, mainly by Ethiopian Telecom, the Ethiopian Science and Technology Commission, Addis Ababa University, the Ge'ez Frontier Foundation, and Ethiopian students abroad. So have, for example, Sisay and Haller (2003) looked at Amharic word formation and lexicon building; Nega and Willett (2002) at stemming; Atelach et al. (2003a) at treebank building; Daniel (Yacob, 2005) at the collection of an (untagged) corpus, tentatively to be hosted by Oxford University's Open Archives Initiative; and Cowell and Hussain (2003) at character recognition.[2] See Atelach et al. (2003b) for an overview of the efforts that have been made so far to develop language processing tools for Amharic.

The need for investigating Amharic information access has been acknowledged by the European Cross-Language Evaluation Forum, which added an Amharic–English track in 2004. However, the task addressed was for accessing an English database in English, with only the original questions being posed in Amharic (and then translated into English). Three groups participated in this track, with Atelach et al. (2004) reporting the best results.

In the present paper we look at the problem of mapping questions posed in Amharic onto a collection of Amharic news items. We use the Self-Organizing Map (SOM) model of artificial neural networks for the task of retrieving the documents matching a specific query. The SOMs were implemented using the Matlab Neural Network Toolbox.

The rest of the paper is laid out as follows. Section 2 discusses artificial neural networks and in particular the SOM model and its application to information access. In Section 3 we describe the Amharic language and its writing system in more detail together with the news items corpora used for training and testing of the networks, while Sections 4 and 5 detail the actual experiments, on text retrieval and text classification, respectively. Finally, Section 6 sums up the main contents of the paper.

2 Artificial Neural Networks

Artificial Neural Networks (ANN) is a computational paradigm inspired by the neurological structure of the human brain, and ANN terminology borrows from neurology: the brain consists of millions of neurons connected to each other through long and thin strands called axons; the connecting points between neurons are called synapses.

ANNs have proved themselves useful in deriving meaning from complicated or imprecise data; they can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computational and statistical techniques. Traditionally, the most common ANN setup has been the backpropagation architecture (Rumelhart et al., 1986), a supervised learning strategy where input data is fed forward in the network to the output nodes (normally with an intermediate hidden layer of nodes) while errors in matches are propagated backwards in the net during training.

2.1 Self-Organizing Maps

Self-Organizing Maps (SOM) is an unsupervised learning scheme neural network, which was invented by Kohonen (1999). It was originally developed to project multi-dimensional vectors on a reduced dimensional space. Self-organizing systems can have many kinds of structures, a common one consists of an input layer and an output layer, with feed-forward connections from input to output layers and full connectivity (connections between all neurons) in the output layer.

A SOM is provided with a set of rules of a local nature (a signal affects neurons in the immediate vicinity of the current neuron), enabling it to learn to compute an input-output pairing with specific desirable properties. The learning process consists of repeatedly modifying the synaptic weights of the connections in the system in response to input (activation) patterns and in accordance to prescribed rules, until a final configuration develops. Commonly both the weights of the neuron closest matching the inputs and the weights of its neighbourhood nodes are increased. At the beginning of the training the neighbourhood (where input patterns cluster depending on their similarity) can be fairly large and then be allowed to decrease over time.

[2] In the text we follow the Ethiopian practice of referring to Ethiopians by their given names. However, the reference list follows Western standard and is ordered according to surnames (i.e., the father's name for an Ethiopian).

2.2 Neural network-based text classification

Neural networks have been widely used in text classification, where they can be given terms and having the output nodes represent categories. Ruiz and Srinivasan (1999) utilize an hierarchical array of backpropagation neural networks for (nonlinear) classification of MEDLINE records, while Ng et al. (1997) use the simplest (and linear) type of ANN classifier, the perceptron. Nonlinear methods have not been shown to add any performance to linear ones for text categorization (Sebastiani, 2002).

SOMs have been used for information access since the beginning of the 90s (Lin et al., 1991). A SOM may show how documents with similar features cluster together by projecting the N-dimensional vector space onto a two-dimensional grid. The radius of neighbouring nodes may be varied to include documents that are weaker related. The most elaborate experiments of using SOMs for document classification have been undertaken using the WEBSOM architecture developed at Helsinki University of Technology (Honkela et al., 1997; Kohonen et al., 2000). WEBSOM is based on a hierarchical two-level SOM structure, with the first level forming histogram clusters of words. The second level is used to reduce the sensitivity of the histogram to small variations in document content and performs further clustering to display the document pattern space.

A Self-Organizing Map is capable of simulating new data sets without the need of retraining itself when the database is updated; something which is not true for Latent Semantic Indexing, LSI (Deerwester et al., 1990). Moreover, LSI consumes ample time in calculating similarities of new queries against all documents, but a SOM only needs to calculate similarities versus some representative subset of old input data and can then map new input straight onto the most similar models without having to recompute the whole mapping.

The SOM model preparation passes through the processes undertaken by the LSI model and the classical vector space model (Salton and McGill, 1983). Hence those models can be taken as particular cases of the SOM, when the neighbourhood diameter is maximized. For instance, one can calculate the LSI model's similarity measure of documents versus queries by varying the SOM's neighbourhood diameter, if the training set is a singular value decomposition reduced vector space. Tambouratzis et al. (2003) use SOMs for categorizing texts according to register and author style and show that the results are equivalent to those generated by statistical methods.

3 Processing Amharic

Ethiopia with some 70 million inhabitants is the third most populous African country and harbours more than 80 different languages.[3] Three of these are dominant: Oromo, a Cushitic language spoken in the South and Central parts of the country and written using the Latin alphabet; Tigrinya, spoken in the North and in neighbouring Eritrea; and Amharic, spoken in most parts of the country, but predominantly in the Eastern, Western, and Central regions. Both Amharic and Tigrinya are Semitic and about as close as are Spanish and Portuguese (Bloor, 1995),

3.1 The Amharic language and script

Already a census from 1994[4] estimated Amharic to be mother tongue of more than 17 million people, with at least an additional 5 million second language speakers. It is today probably the second largest language in Ethiopia (after Oromo). The Constitution of 1994 divided Ethiopia into nine fairly independent regions, each with its own nationality language. However, Amharic is the language for countrywide communication and was also for a long period the principal literal language and medium of instruction in primary and secondary schools in the country, while higher education is carried out in English.

Amharic and Tigrinya speakers are mainly Orthodox Christians, with the languages drawing common roots to the ecclesiastic Ge'ez still used by the Coptic Church. Both languages are written using the Ge'ez script, horizontally and left-to-right (in contrast to many other Semitic languages). Written Ge'ez can be traced back to at least the 4th century A.D. The first versions of the script included consonants only, while the characters in later versions represent consonant-vowel (CV) phoneme pairs. In modern written Amharic, each syllable pat-

[3] How many languages there are in a country is as much a political as a linguistic issue. The number of languages of Ethiopia and Eritrea together thus differs from 70 up to 420, depending on the source; however, 82 (plus 4 extinct) is a common number.

[4] Published by Ethiopia's Central Statistal Authority 1998.

Order	1	2	3	4	5	6	7
C \ V	/ə/	/u/	/i/	/ɐ/	/e/	/ɨ/	/o/
/s/	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
/m/	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ

Table 1: The orders for ስ (/s/) and ም (/m/)

tern comes in seven different forms (called *orders*), reflecting the seven vowel sounds. The first order is the basic form; the other orders are derived from it by more or less regular modifications indicating the different vowels. There are 33 basic forms, giving 7*33 syllable patterns, or *fidEls*.

Two of the base forms represent vowels in isolation (ዐ and አ), but the rest are for consonants (or semivowels classed as consonants) and thus correspond to CV pairs, with the first order being the base symbol with no explicit vowel indicator (though a vowel is pronounced: C+/ə/). The sixth order is ambiguous between being just the consonant or C+/ɨ/. The writing system also includes 20 symbols for labialised velars (four five-character orders) and 24 for other labialisation. In total, there are 275 *fidEls*. The sequences in Table 1 (for ስ and ም) exemplify the (partial) symmetry of vowel indicators.

Amharic also has its own numbers (twenty symbols, though not widely used nowadays) and its own punctuation system with eight symbols, where the space between words looks like a colon ፡, while the full stop, comma and semicolon are ።, ፣ and ፤. The question and exclamation marks have recently been included in the writing system. For more thorough discussions of the Ethiopian writing system, see, for example, Bender et al. (1976) and Bloor (1995).

Amharic words have consonantal roots with vowel variation expressing difference in interpretation, making stemming a not-so-useful technique in information retrieval (no full morphological analyser for the language is available yet). There is no agreed upon spelling standard for compounds and the writing system uses multitudes of ways to denote compound words. In addition, not all the letters of the Amharic script are strictly necessary for the pronunciation patterns of the language; some were simply inherited from Ge'ez without having any semantic or phonetic distinction in modern Amharic: there are many cases where numerous symbols are used to denote a single phoneme, as well as words that have extremely different orthographic form and slightly distinct phonetics, but the same meaning. As a result of this, lexical variation and homophony is very common, and obviously deteriorates the effectiveness of Information Access systems based on strict term matching; hence the basic idea of this research: to use the approximative matching enabled by self-organizing map-based artificial neural networks.

3.2 Test data and preprocessing

In our SOM-based experiments, a corpus of news items was used for text classification. A main obstacle to developing applications for a language like Amharic is the scarcity of resources. No large corpora for Amharic exist, but we could use a small corpus of 206 news articles taken from the electronic news archive of the website of the Walta Information Center (an Ethiopian news agency). The training corpus consisted of 101 articles collected by Saba (Amsalu, 2001), while the test corpus consisted of the remaining 105 documents collected by Theodros (GebreMeskel, 2003). The documents were written using the Amharic software VG2 Main font.

The corpus was matched against 25 queries. The selection of documents relevant to a given query, was made by two domain experts (two journalists), one from the Monitor newspaper and the other from the Walta Information Center. A linguist from Gonder College participated in making consensus of the selection of documents made by the two journalists. Only 16 of the 25 queries were judged to have a document relevant to them in the 101 document training corpus. These 16 queries were found to be different enough from each other, in the content they try to address, to help map from document collection to query contents (which were taken as class labels). These mappings (assignment) of documents to 16 distinct classes helped to see retrieval and classification effectiveness of the ANN model.

The corpus was preprocessed to normalize spelling and to filter out stopwords. One preprocessing step tried to solve the problems with non-standardised spelling of compounds, and that the same sound may be represented with two or more distinct but redundant written forms. Due to the systematic redundancy inherited from the Ge'ez, only about 233 of the 275 *fidEls* are actually necessary to

Sound pattern	Matching Amharic characters
/sə/	ሰ, ሠ
/ɾə/	ጸ, ፀ
/hə/	ሀ, ሃ, ሐ, ሓ, ኀ, ኃ
/iə/	አ, ኣ, ዐ, ዓ

Table 2: Examples of character redundancy

represent Amharic. Some examples of character redundancy are shown in Table 2. The different forms were reduced to common representations.

A negative dictionary of 745 words was created, containing both stopwords that are news specific and the Amharic text stopwords collected by Nega (Alemayehu and Willett, 2002). The news specific common terms were manually identified by looking at their frequency. In a second preprocessing step, the stopwords were removed from the word collection before indexing. After the preprocessing, the number of remaining terms in the corpus was 10,363.

4 Text retrieval

In a set of experiments we investigated the development of a retrieval system using Self-Organizing Maps. The term-by-document matrix produced from the entire collection of 206 documents was used to measure the retrieval performance of the system, of which 101 documents were used for training and the remaining for testing. After the preprocessing described in the previous section, a weighted matrix was generated from the original matrix using the log-entropy weighting formula (Dumais, 1991). This helps to enhance the occurrence of a term in representing a particular document and to degrade the occurrence of the term in the document collection. The weighted matrix can then be dimensionally reduced by Singular Value Decomposition, SVD (Berry et al., 1995). SVD makes it possible to map individual terms to the concept space.

A query of variable size is useful for comparison (when similarity measures are used) only if its size is matrix-multiplication-compatible with the documents. The pseudo-query must result from the global weight obtained in weighing the original matrix to be of any use in ranking relevant documents. The experiment was carried out in two versions, with the original vector space and with a reduced one.

4.1 Clustering in unreduced vector space

In the first experiment, the selected documents were indexed using 10,363 dimensional vectors (i.e., one dimension per term in the corpus) weighted using log-entropy weighting techniques. These vectors were fed into an Artificial Neural Network that was created using a SOM lattice structure for mapping on a two-dimensional grid. Thereafter a query and 101 documents were fed into the ANN to see how documents cluster around the query.

For the original, unnormalised (unreduced, 10,363 dimension) vector space we did not try to train an ANN model for more than 5,000 epochs (which takes weeks), given that the network performance in any case was very bad, and that the network for the reduced vector space had its apex at that point (as discussed below).

Those documents on the node on which the single query lies and those documents in the immediate vicinity of it were taken as being relevant to the query (the neighbourhood was defined to be six nodes). Ranking of documents was performed using the cosine similarity measure, on the single query versus automatically retrieved relevant documents. The eleven-point average precision was calculated over all queries. For this system the average precision on the test set turned out to be 10.5%, as can be seen in the second column of Table 3.

The table compares the results on training on the original vector space to the very much improved ones obtained by the ANN model trained on the reduced vector space, described in the next section.

Recall	Original vector	Reduced vector
0.00	0.2080	0.8311
0.10	0.1986	0.7621
0.20	0.1896	0.7420
0.30	0.1728	0.7010
0.40	0.0991	0.6888
0.50	0.0790	0.6546
0.60	0.0678	0.5939
0.70	0.0543	0.5300
0.80	0.0403	0.4789
0.90	0.0340	0.3440
1.00	0.0141	0.2710
Average	0.1052	0.5998

Table 3: Eleven-point precision for 16 queries

4.2 Clustering in SVD-reduced vector space

In a second experiment, vectors of numerically indexed documents were converted to weighted matrices and further reduced using SVD, to infer the need for representing co-occurrence of words in identifying a document. The reduced vector space of 101 pseudo-documents was fed into the neural net for training. Then, a query together with 105 documents was given to the trained neural net for simulation and inference purpose.

For the reduced vectors a wider range of values could be tried. Thus 100, 200, ..., 1000 epochs were tried at the beginning of the experiment. The network performance kept improving and the training was then allowed to go on for 2000, 3000, ..., 10,000, 20,000 epochs thereafter. The average classification accuracy was at an apex after 5,000 epochs, as can been seen in Figure 1.

The neural net with the highest accuracy was selected for further analysis. As in the previous model, documents in the vicinity of the query were ranked using the cosine similarity measure and the precision on the test set is illustrated in the third column of Table 3. As can be seen in the table, this system was effective with 60.0% eleven-point average precision on the test set (each of the 16 queries was tested).

Thus, the performance of the reduced vector space system was very much better than that obtained using the test set of the normal term document matrix that resulted in only 10.5% average precision. In both cases, the precision of the training set was assessed using the classification accuracy which shows how documents with similar features cluster together (occur on the same or neighbouring nodes).

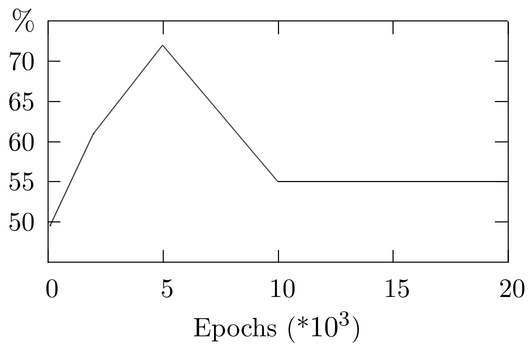

Figure 1: Average network classification accuracy

5 Document Classification

In a third experiment, the SVD-reduced vector space of pseudo-documents was assigned a class label (query content) to which the documents of the training set were identified to be more similar (by experts) and the neural net was trained using the pseudo-documents and their target classes. This was performed for 100 to 20,000 epochs and the neural net with best accuracy was considered for testing.

The average precision on the training set was found to be 72.8%, while the performance of the neural net on the test set was 69.5%. A matrix of simple queries merged with the 101 documents (that had been used for training) was taken as input to a SOM-model neural net and eventually, the 101-dimensional document and single query pairs were mapped and plotted onto a two-dimensional space. Figure 2 gives a flavour of the document clustering.

The results of this experiment are compatible with those of Theodros (GebreMeskel, 2003) who used the standard vector space model and latent semantic indexing for text categorization. He reports that the vector space model gave a precision of 69.1% on the training set. LSI improved the precision to 71.6%, which still is somewhat lower than the 72.8% obtained by the SOM model in our experiments. Going outside Amharic, the results can be compared to the ones reported by Cai and Hofmann (2003) on the Reuters-21578 corpus[5] which contains 21,578 classified documents (100 times the documents available for Amharic). Used an LSI approach they obtained document average precision figures of 88–90%.

In order to locate the error sources in our experiments, the documents missed by the SOM-based classifier (documents that were supposed to be clustered on a given class label, but were not found under that label), were examined. The documents that were rejected as irrelevant by the ANN using reduced dimension vector space were found to contain only a line or two of interest to the query (for the training set as well as for the test set). Also within the test set as well as in the training set some relevant documents had been missed for unclear reasons.

Those documents that had been retrieved as relevant to a query without actually having any relevance to that query had some words that co-occur

[5] Available at `www.daviddlewis.com/resources`

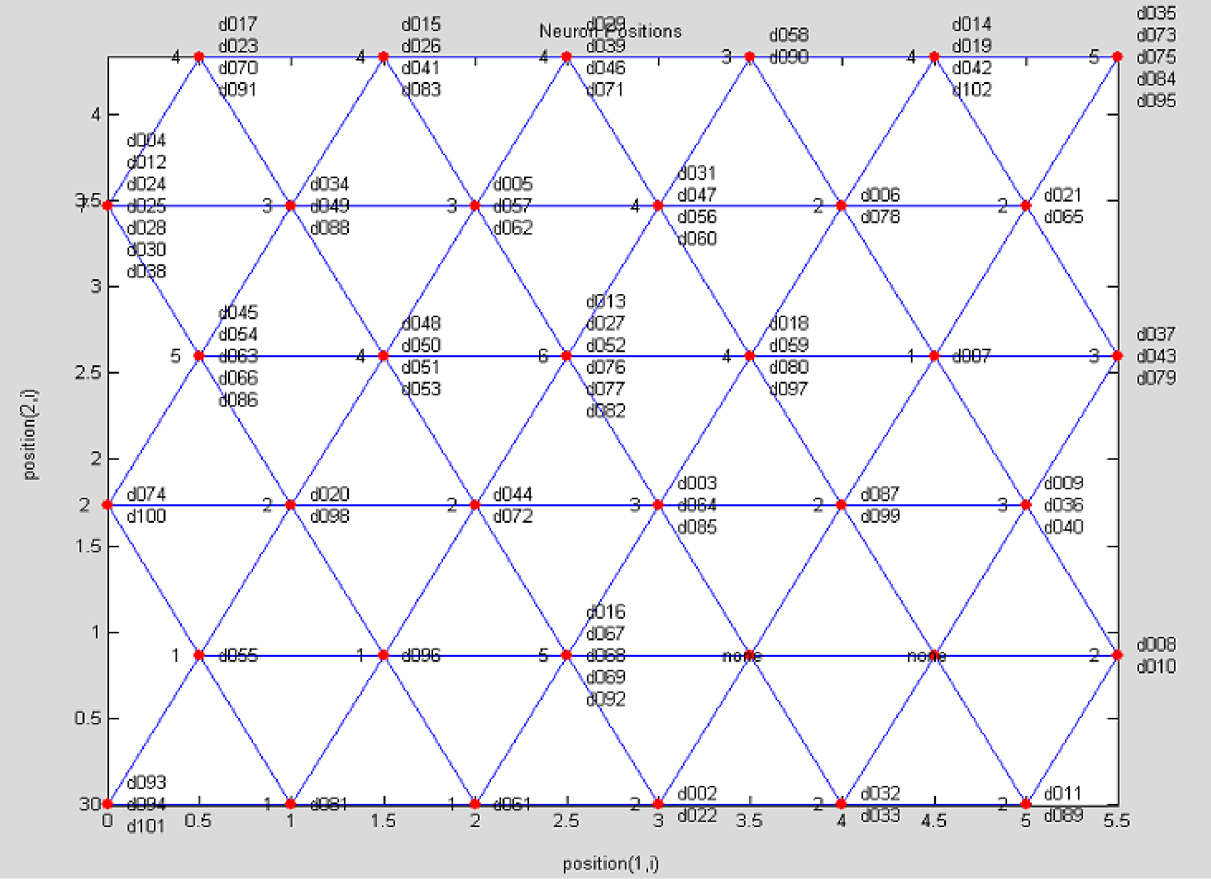

Figure 2: Document clustering at different neuron positions

with the words of the relevant documents. Very important in this observation was that documents that could be of some interest to two classes were found at nodes that are the intersection of the nodes containing the document sets of the two classes.

6 Summary and Conclusions

A set of experiments investigated text retrieval of selected Amharic news items using Self-Organizing Maps, an unsupervised learning neural network method. 101 training set items, 25 queries, and 105 test set items were selected. The content of each news item was taken as the basis for document indexing, and the content of the specific query was taken for query indexing. A term–document matrix was generated and the occurrence of terms per document was registered. This original matrix was changed to a weighted matrix using the log-entropy scheme. The weighted matrix was further reduced using SVD. The length of the query vector was also reduced using the global weight vector obtained in weighing the original matrix.

The ANN model using unnormalised vector space had a precision of 10.5%, whereas the best ANN model using reduced dimensional vector space performed at a 60.0% level for the test set. For this configuration we also tried to classify the data around a query content, taken that query as class label. The results obtained then were 72.8% for the training set and 69.5% for the test set, which is encouraging.

7 Acknowledgments

Thanks to Dr. Gashaw Kebede, Kibur Lisanu, Lars Asker, Lemma Nigussie, and Mesfin Getachew; and to Atelach Alemu for spotting some nasty bugs.

The work was partially funded by the Faculty of Informatics at Addis Ababa University and the ICT support programme of SAREC, the Department for Research Cooperation at Sida, the Swedish International Development Cooperation Agency.

References

Nega Alemayehu and Peter Willett. 2002. Stemming of Amharic words for information retrieval. *Literary and Linguistic Computing*, 17(1):1–17.

Atelach Alemu, Lars Asker, and Gunnar Eriksson. 2003a. An empirical approach to building an Amharic treebank. In *Proc. 2nd Workshop on Treebanks and Linguistic Theories*, Växjö University, Sweden.

Atelach Alemu, Lars Asker, and Mesfin Getachew. 2003b. Natural language processing for Amharic: Overview and suggestions for a way forward. In *Proc. 10th Conf. Traitement Automatique des Langues Naturelles*, Batz-sur-Mer, France, pp. 173–182.

Atelach Alemu, Lars Asker, Rickard Cöster, and Jussi Karlgren. 2004. Dictionary-based Amharic–English information retrieval. In *5th Workshop of the Cross Language Evaluation Forum*, Bath, England.

Saba Amsalu. 2001. The application of information retrieval techniques to Amharic. MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia.

Marvin Bender, Sydney Head, and Roger Cowley. 1976. The Ethiopian writing system. In Bender et al., eds, *Language in Ethiopia*. Oxford University Press.

Michael Berry, Susan Dumais, and Gawin O'Brien. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review*, 37(4):573–595.

Thomas Bloor. 1995. The Ethiopic writing system: a profile. *Journal of the Simplified Spelling Society*, 19:30–36.

Lijuan Cai and Thomas Hofmann. 2003. Text categorization by boosting automatically extracted concepts. In *Proc. 26th Int. Conf. Research and Development in Information Retrieval*, pp. 182–189, Toronto, Canada.

John Cowell and Fiaz Hussain. 2003. Amharic character recognition using a fast signature based algorithm. In *Proc. 7th Int. Conf. Image Visualization*, pp. 384–389, London, England.

Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Susan Dumais. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236.

Sisay Fissaha and Johann Haller. 2003. Application of corpus-based techniques to Amharic texts. In *Proc. MT Summit IX Workshop on Machine Translation for Semitic Languages*, New Orleans, Louisana.

Jane Furzey. 1996. Enpowering socio-economic development in Africa utilizing information technology. A country study for the United Nations Economic Commission for Africa, University of Pennsylvania.

Theodros GebreMeskel. 2003. Amharic text retrieval: An experiment using latent semantic indexing (LSI) with singular value decomposition (SVD). MSc Thesis, School of Information Studies for Africa, Addis Ababa University, Ethiopia.

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen. 1997. WEBSOM — Self-Organizing Maps of document collections. In *Proc. Workshop on Self-Organizing Maps*, pp. 310–315, Espoo, Finland.

Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, and Antti Saarela. 2000. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585.

Teuvo Kohonen. 1999. *Self-Organization and Associative Memory*. Springer, 3 edition.

Xia Lin, Dagobert Soergel, and Gary Marchionini. 1991. A self-organizing semantic map for information retrieval. In *Proc. 14th Int. Conf. Research and Development in Information Retrieval*, pp. 262–269, Chicago, Illinois.

Hwee Tou Ng, Wei Boon Goh, and Kok Leong Low. 1997. Feature selection, perceptron learning, and a usability case study for text categorization. In *Proc. 20th Int. Conf. Research and Development in Information Retrieval*, pp. 67–73, Philadelphia, Pennsylvania.

Miguel Ruiz and Padmini Srinivasan. 1999. Hierarchical neural networks for text categorization. In *Proc. 22nd Int. Conf. Research and Development in Information Retrieval*, pp. 281–282, Berkeley, California.

David Rumelhart, Geoffrey Hinton, and Ronald Williams. 1986. Learning internal representations by error propagation. In Rumelhart and McClelland, eds, *Parallel Distributed Processing*, vol 1. MIT Press.

Gerard Salton and Michael McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

George Tambouratzis, N. Hairetakis, S. Markantonatou, and G. Carayannis. 2003. Applying the SOM model to text classification according to register and stylistic content. *Int. Journal of Neural Systems*, 13(1):1–11.

Daniel Yacob. 2005. Developments towards an electronic Amharic corpus. In *Proc. TALN 12 Workshop on NLP for Under-Resourced Languages*, Dourdan, France, June (to appear).