

# Applying Machine Learning to Amharic Text Classification

Lars Asker<sup>1</sup>, Atelach Alemu Argaw<sup>1</sup>, Björn Gambäck<sup>2</sup> and Magnus Sahlgren<sup>2</sup>

Stockholm University<sup>1</sup> and Swedish Institute of Computer Science<sup>2</sup>

## 1. Introduction

The goal of this work has been to investigate how well a high-level task like text classification can be carried out on Amharic text. We have also investigated the effect that operations like stemming or part-of-speech tagging can have on text classification performance for such a highly inflectional language like Amharic.

Amharic is the official working language of the federal government of the Federal Democratic Republic of Ethiopia and spoken by well over 20 million people as a first or second language. The actual size of the Amharic speaking population has to be based on estimates: Hudson (1999) analyzed the latest national Ethiopian census from 1994 and indicated that more than 40% of the (then) 53 million Ethiopians understood Amharic, with at the time about 17 million first language speakers. The current size of the Ethiopian population is estimated to be some 75 million people (CIA 2006). Amharic is the second most spoken Semitic language in the world (after Arabic) and closely related to Tigrinya. It is today probably the second largest language in Ethiopia (after Oromo, a Cushitic language) and possibly one of the five largest languages on the African continent. Following the Constitution drafted in 1993, Ethiopia is divided into nine fairly independent regions, each with its own nationality language. However, Amharic is the language for country-wide communication and was also for a long period the principal literal language and medium of instruction in primary and secondary schools of the country, while higher education is carried out in English.

In spite of the relatively large number of speakers, Amharic is still a language for which very few computational linguistic resources have been developed, and very little has been done in terms of making useful higher level Internet or computer based applications available to those who only speak Amharic. It is generally believed that applications such as information retrieval, text classification, or document filtering could benefit from the existence and availability of basic tools such as stemmers, morphological analyzers or part-of-speech taggers. However, since so few language processing resources for Amharic are available, very little is known about their effect on retrieval or classification performance for this language. It has been argued that stemming can improve text categorization performance, especially for highly inflected languages like Amharic, and it has therefore been the goal of this work to explore this issue further.

## 2. Stemming

Stemming is a technique whereby morphological variants are reduced to a single stem. It is language dependent and should be tailored for each language since languages have a varying degree of differences in their morphological properties. It should also be tailored to the specific task at hand. Stemming is commonly applied in text classification tasks as a pre-processing method to reduce morphological variants to a single feature. Stop word removal is another technique used in the preprocessing stage. Here the purpose is to remove commonly occurring words e.g. “the”, “on”, “he” etc since they do not help in discriminating between documents.

Stemming for information retrieval and text classification tasks has been reported to show mixed results. For example Gaustad and Bouma (2002) report results from experiments on Dutch email and news text classification using simple suffix stripping and a dictionary based stemming. They report that neither improve classification accuracy. Classification and retrieval tasks for English commonly apply stemming in the pre-processing stage, although the effects on the performance are not conclusive. On the other hand, similar research on highly inflected languages such as Arabic report an increase in performance due to stemming for classification tasks (cf. Syiam et al. 2006). The effect of stemming in automatic classification tasks is not consistent and is claimed to be dependent on the specific language and the domain of the document collection. Amharic is a language with very rich morphology; hence it is intuitively assumed that stemming will have a positive effect for classification and related tasks.

## 3. Related Work

The main previous contributions in the area include the work by Nega and Willett (2002, 2003) in which they investigated the effect of stemming for information retrieval on a limited Amharic document collection (consisting of 548 documents and 40 queries). Habte (2006) carried out some small experiments on classification of documents into 10 news categories (the same categories as used in this paper, see below), using an artificial neural network approach (Self-Organizing Maps, SOM) on a corpus of 100 news items, while Eyassu and Gambäck (2005) reported on experiments on classifying news items according to a set of queries, taking the queries as class labels. Their experiments were based on a 206 document corpus which was converted to a term-document matrix, reduced using Singular Value Decomposition (a mathematical operation to reduce dimensionality) and also passed through SOM.

Common to all the above-mentioned previous efforts has been that they have been severely hampered by the lack of large-scale linguistic resources for Amharic. The work by Argaw and Asker (2006) has been trying to remedy this by accessing machine-readable dictionaries. In the present paper we utilize a medium-sized, hand-tagged Amharic corpus from Addis Ababa University (Demeke and Getachew 2006) and investigate what effect this information combined with stemming has on the task of classifying news items according to categories.

## 4. Amharic

Written Amharic (and Tigrinya) uses a unique script which has originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). Written Ge'ez can be traced back to at least the 4th century A.D. The first versions of the language included consonants only, while the characters in later versions represent consonant-vowel (CV) phoneme pairs. In the modern Ethiopic script each syllable pattern 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  $7 \times 33$  syllable patterns (syllographs), or *fidEls*. Two of the base forms represent vowels in isolation, but the rest are for consonants (or semi-vowels classed as consonants) and thus correspond to CV pairs, with the first order being the base symbol with no explicit vowel indicator. The writing system also includes four (incomplete, five-character) orders of labialised velars and 24 additional labialised consonants. In total, there are 275 *fidEls*. The script also has a unique set of punctuation marks and digits. Unlike Arabic or Hebrew, the language is written from left to right. Like many other Semitic languages, Amharic has a rich verb morphology which is based on triconsonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. The Amharic writing system uses multitudes of ways to denote compound words and there is no agreed upon spelling standard for compounds. As a result of this - and of the size of the country leading to vast dialectal dispersion - lexical variation and homophony is very common. In addition, not all the letters of the Amharic script are strictly necessary for the pronunciation patterns of the spoken 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 with the same meaning. So are, for example, most labialised consonants basically redundant, and there are actually only 39 context-independent phonemes (monophones): of the 275 symbols of the script, only about 233 remain if the redundant ones are removed.

## 5. The Corpus

In the experiments described here we have used a corpus consisting of 1065 Amharic news texts from the years 2001 – 2002 (1994 according to the Ethiopian calendar) that were collected from the Walta Information Center<sup>1</sup>. The corpus contains a total of approximately 210.000 words and has been made available in XML-format in both Amharic script and in a transliterated form.

---

<sup>1</sup> Walta Information Center is a private news and information service located in Addis Ababa, Ethiopia. At its web site [www.waltainfo.com](http://www.waltainfo.com), it provides Ethiopia related news in English and Amharic on a daily basis.

The original Amharic text was transliterated into SERA (System for Ethiopic Representation in ASCII) using a file conversion utility called g2<sup>ii</sup> which is available in the LibEth package<sup>iii</sup>. We worked with the transliterated form in order to make it compatible with the machine learning toolbox used for the experiments. Figure 1 below shows an example of one of the news texts.

```
<filename> mes07a2.htm </filename>
<title>
<fidel>
በቦረና የኦህዴድ ተሃድሶ ውይይት ተጀመረ
</fidel>
<sera>
beborena yeohdEd tehadso wyyt tejemere
</sera>
</title>
<dateline place="negele" month="meskerem" date="7/1994/(WIC)"/>
<body>
<fidel> በቦረና ዞንና 13 ወረዳዎች ለሚገኙ የመንግሥት ሠራተኞች የተዘጋጀ የኦህዴድ ተሃድሶ ውይይት ዛሬ መጀመሩን የዙ መስተዳድር ምክር ቤት አስታወቀ። የምክር ቤቱ ፀሐፊ አቶ መሐመድ ጅሎ እንደገለፁት ለአምስት ቀናት በሚቆየው በዚህ ተሃድሶ የአብዮታዊ ዲሞክራሲያዊ ጥያቄ በኢትዮጵያ ፣ የአብዮታዊ ዲሞክራሲ የልማት መርሆዎች ፣ ስትራቴጂዎችና የሥርዓቱ አደጋዎች በሚሉ ርዕሶች ላይ ውይይት ይካሄዳል። የአመለካከትን ጥራት ለማምጣት በሚካሄደው የተሃድሶ ውይይት ከዞን መምሪያዎችና ከወረዳ ጽሕፈት ቤቶች የተውጣጡ ከ2 ሺ 500 በላይ የመንግሥት ሠራተኞች ይሳተፋሉ ተብሎ እንደሚጠበቅ ፀሐፊው ለዋልታ ኢንፎርሜሽን ማዕከል ገልፀዋል። </fidel>
<sera> beborena zonna 13 weredawoc lemigeNu yemeng`st `serateNoc yetezegaje yeohdEd tehadso wyyt zarE mejemerun yezonu mestedadr mkr bEt astaweqe:: yemkr bEtu `SeHefi ato meHemed jlo Indegele`Sut leamst qenat bemiqoyew bezihu tehadso yeabyotawi dEmokrasiyawi TyaqE beityoPya, yeabyotawi dEmokrasi yelmat merhowoc, stratEjjiwocna ye`s`atu adegawoc bemilu r`Isoc lay wyyt ykahEdal:: yeamelekaketrn Trat lemmTat bemikahEdew yetehadso wyyt kezon memriyawocna kewereda SHfet bEtoC yetewTaTu ke2 xi 500 belay yemeng`st `serateNoc ysatefalu teblo IndemiTebeq `SeHefiw lewalta informExn ma`Ikel gel`Sewal:: </sera>
<copyright> Copyright 1998 - 2002 Walta Information Center </copyright>
</body>
</document>
```

Figure 1. An example of one of the news texts.

The corpus has also been morphologically analyzed and POS-tagged and each article has been manually classified as belonging to one of 10 predefined classes. The ten classes are presented in Figure 2 below.

<sup>ii</sup> g2 was made available to us by Daniel Yacob of the Ge'ez Frontier Foundation (<http://www.ethiopic.org/>)  
<sup>iii</sup> LibEth is a library for Ethiopic text processing written in ANSI C (<http://libeth.sourceforge.net/>)

Cat	Topic	#
0	Sport	9
1	Hot News	55
2	Editorials	0
3	Politics	140
4	Business & Economy	351
5	Social	356
6	Culture	11
7	Science & Technology	47
8	Health	93
9	Art	3
		1065

Figure 2. The ten categories and the number of articles belonging to each category.

## 6. The Representation

For the text classification experiment, we used a bag-of-words approach represented as a vector-space model. That is, each article in the corpus was represented as a (sparse) binary vector and where each position in the vector corresponds to a specific unique word that occurs in at least one of the news articles in the corpus. If an article contains a word, then the vector position corresponding to that word would have the value 1, else it would be 0. Hence, each article is represented as a vector, by a sequence of ones and zeros, where the ones correspond to those specific words that occur in the article.

Since the idea is to investigate how stemming affects text categorization performance, and also to compare it with a representation based on only the nouns that occur in the text, we used three different representations for each article. The first representation used the full text for each article and represented it in the form of a binary vector as described above. In the rest of the paper, this representation will be referred to as "**full**". The second representation instead used a stemmed version of the text to represent each document. For this, we developed a morphological analyzer and stemmer for Amharic. The morphological analyzer finds all possible segmentations of a given word according to the morphological rules of the language and then selects the most likely prefix and suffix for the word based on corpus statistics. It strips off the prefix and suffix and then tries to look up the remaining stem (or alternatively, some morphologically motivated variants of it) in a dictionary to verify that it is a possible stem of the word. The frequency and distribution of prefixes and suffixes over Amharic words is based on a statistical analysis of a 3.5 million word Amharic news corpus. The morphological analyzer/stemmer had an accuracy of approximately 85% when evaluated on a limited text consisting of 50 sentences (805 words) from

this years CLEF (Argaw and Asker 2006). Figure 3 below shows some examples of words and their stemmed form. In this context it is interesting to note that incorrectly stemmed words will have no negative effect on classification performance so long as other words that belong to a different category are not reduced to the exact same form. See e.g. the incorrect form ezo in Figure 3, it works just as well as the correct form zon as a stemmed form for the words yezon, yezonu, and bezonu.

Stemmed	Non Stemmed
-----	
br	br bebr
bexta	bexta bextaw bextawoc bebextaw kebextaw
mkr	mkr yemkr bemkr
xeT	yetexeTebet texeTe mexeTun
mekelakeya	mekelakeyana mekelakeya yemekelakeya
mrCa	mrCa yemrCa
mnzari	mnzari yemnzari
guba	gubaE gubaEw
bEt	bEt bEtu yebEt bEtoc bEtocn bEtocna bEte
projekt	projektoc projectocu projectocn projekt yeprojekt yeprojektu projektu
bank	bank bebankoc banku bankoc yebank
guday	guday gudayu gudayoc
zqteNa	zqteNaw zqteNa
amakay	beamakay amakaynet
ezo	yezon yezonu bezonu
mrmr	mrmr bemrmr yemrmr
hzb	hzb hzbu hzboc hzb yehzb lehzb
lmat	lmat yelmat lelmat lmatn lmatna yelmatna belmat
Ec	Ec yeEc beEc

Figure 3. Stemmed form and original variants of words occurring in the text

Since Amharic, just like other Semitic languages, has a very rich morphology<sup>iv</sup>, it was expected that the stemming would reduce the size of the representation considerably and also to improve classification accuracy. The second representation was thus to use the stemmed version of each article and represent it in the form of a binary vector as described above. In the rest of the paper, this representation will be referred to as "**stemmed**".

In the third representation we used only the nouns from the full text of each article and represent it in the form of a binary vector as described above. In the rest of the paper, this representation will be referred to as "**nouns**".

The corpus has been manually POS-tagged by staff at the Ethiopian Languages Research Center at Addis Ababa University<sup>v</sup> (Demeke and Getachew 2006). The tagset consists of 10 basic

<sup>iv</sup> A verb could for example have well over 150 different forms.

<sup>v</sup> <http://www.aau.edu.et/research/elrc/>

classes: Noun, Pronoun, Verb, Adjective, Preposition, Conjunction, Adverb, Numeral, Interjection, and Punctuation plus one extra tag for problematic (unclassified) words. The 10 basic classes were then further divided into a total of 30 subclasses.

In the third representation ("**nouns**"), we only used those words in each news article that had been labeled as nouns by the POS-tagging. The reasoning behind this is that nouns tend to carry more information than any other word classes. For example Hulth (2004) investigated the frequencies of different POS patterns in keywords that have been assigned to documents by professional indexers, and have found that as many as 90% of keywords consist of nouns and noun phrases. We therefore investigated to which extent it was possible to use only the nouns to represent the text without losing too much in classification performance compared to that of the other representations.

## 7. The Experiment

The experiments were set up and run using the RDS system<sup>vi</sup>. RDS is a rule based machine learning platform that supports a variety of rule based induction techniques such as rule sets and decision trees, and ensemble methods like bagging, boosting and random forests. In the experiments, we compared the three representations using 10-fold cross validation while constructing classification models based on bagging of decision trees.

A decision tree is a predictive model that can be automatically constructed from labeled training examples by initially partitioning the examples into two or more groups according to the values of the most discriminative attribute. The decision tree construction is then continued by recursively partitioning each sub-group in the same way until all examples in each sub-group have the same label and no further partitioning is possible. Once constructed, such a decision tree can then be used to predict the category of new and unseen examples.

Bagging (bootstrap aggregation) is a machine learning technique used to improve the performance of a combined classifier by training and combining the predictions of several (usually 20 – 50) different component classifiers. The component classifiers vary by being trained on different variants of the original training data set. The variants are constructed by randomly, and with replacement, selecting N examples from the original training data set, where N is the total number of examples in the original training set.

10-fold cross validation is a statistical technique used to estimate the performance of e.g. a machine learning classifier. It involves partitioning a data set into ten non-overlapping subsets and then using each subset for testing once, while the other nine are used for training of the classifier. The whole process is repeated ten times and the results for each are combined to get an overall estimate of the classifier. The results of the experiments are presented in Figure 4, below.

---

<sup>vi</sup> Rule Discovery System (RDS) is available from Compumine AB (<http://www.compumine.com>).

Method	Accuracy
stemmed	68.08 %
nouns	68.92 %
full	69.39 %

Figure 4. Results from the experiments.

## 8. Conclusion

It has been claimed that stemming is an important preprocessing step that will allow for improved text categorization accuracy, especially for languages such as Amharic that has a rich inflectional morphology. In order to investigate this further, we have performed an experiment where the goal has been to see how stemming vs. non-stemming will affect the performance in a text categorization task. The best accuracy (69.39%) was achieved using the full text as representation. The representation using only the nouns performed almost equally well, confirming the assumption that most of the information required for distinguishing between the various categories actually is contained in the nouns.

The classifier using the stemmed representation only performs *almost* on par with the other classifiers. It might seem a bit surprising that the stemmed representation failed to improve the performance of the text classifier on such a morphologically rich language as Amharic. An explanation for the lack of improvement might be that the main morphological variation in Amharic pertains to the verbs, not the nouns. However, the latter are the main carriers of the information content of the language, and hence the main sources of information for the classifiers, which is also partly confirmed by the results of the experiments. Further experiments are however still needed in order to get a better understanding of the factors that influence text categorization performance for Amharic.

## References

- Alemayehu, Nega, and Peter Willett. 2002. Stemming of Amharic Words for Information Retrieval. *Literary and Linguistic Computing* 17(1).1-17.
- Alemayehu, Nega, and Peter Willett. 2003. The effectiveness of stemming for information retrieval in Amharic. *Emerald Research Register* 37(4).254-259.
- Argaw, Atelach Alemu, and Lars Asker. 2006. Amharic-English Information Retrieval. Working Notes of Cross Language Evaluation Forum (CLEF).

CIA. 2006. The World Factbook - Ethiopia. The Central Intelligence Agency. Washington, DC. (latest update: 19 December, 2006).

Demeke, Girma A., and Mesfin Getachew. 2006. Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges. ELRC Working Papers, Vol II, Number 1.

Eyassu, Samuel, and Björn Gambäck. 2005. Classifying Amharic News Text Using Self-Organizing Maps. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics Workshop on Computational Approaches to Semitic Languages.

Gaustad, Tanja and Gosse Bouma. 2002. Accurate Stemming of Dutch for Text Classification. Computational Linguistics in the Netherlands.

Habte, Lemma Nigussie. 2006. Collaborative News Filtering for Amharic: An Experiment Using Neural Networks. Master of Science Thesis in Information Science, Addis Ababa University.

Hudson, Grover. 1999. Linguistic analysis of the 1994 Ethiopian Census. Northeast African Studies.6(3).89-107. Michigan University Press.

Hulth, Anette. 2004. Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction. Doctoral Dissertation, Department of Computer and Systems Sciences, Stockholm University.

Syiam, M. M., Z. T. Fayed, and M. B. Habib. 2006. An Intelligent System for Arabic Text Classification. IJICIS 6(1).