WEB MINING FOR AN AMHARIC - ENGLISH BILINGUAL CORPUS

Atelach Alemu Argaw and Lars Asker
Department of Computer and Systems Sciences
Stockholm University
Forum 100
SE-164 40 Kista, Sweden

Keywords: Web mining, parallel corpora, text alignment, Amharic.

Abstract: We present recent work aimed at constructing a bilingual corpus consisting of comparable Amharic and English news texts. The Amharic and English texts were collected from an Ethiopian news agency that publishes daily news in Amharic and English through their web page. The Amharic texts are represented using Ethiopic script and archived according to the Ethiopian calender. The overlap between the corresponding Amharic and English news texts in the archive is comparatively small, only approximately one article out of ten has a corresponding translated version. Thus a major part of the work has been to identify the subset of matching news texts in the archive, transliterating the Amharic texts into an ASCII representation, and aligning them with their respective corresponding English version. In doing so, we utilised a number of available software and data sources that were (mainly) found on the Internet. Amharic is a language for which very few computational linguistic tools or corpora (such as electronic lexica, part-of-speech taggers, parsers or tree-banks) exist. A challenge has therefor been to show that it is possible to create a comparable corpus even in the absence to these resources. We used fuzzy string matching between words in the English and Amharic titles as a way to determine how likely it is that two news items are referring to the same event. In order to restrict the matching algorithm further, we only compared titles of news items that were published on the corresponding same date and at the same place. We present an experimental evaluation of the algorithm, based on data from one year, and show that fuzzy string matching of news titles can be sufficient to align Amharic and English news text with relatively high precision despite the obvious difference between the two languages.

1 INTRODUCTION

The emergence of the World Wide Web has provided new and important opportunities to easily access and combine data and information from several different sources and thereby enabling the construction of new resources for researchers and people everywhere. Although most web pages on the Internet are in English, there is a growing number of non-English web pages, and today there are more non-English than English speaking Internet users (GlobalReach, 2004).

For people in many parts of the world there is an urgent need to develop tools and resources that can allow them to better access, process and disseminate information on the Internet while using their own language. One way to develop these resources is through the transfer of linguistic knowledge, tools and techniques from languages where these resources have been more developed and are more readily available. An important requisite for this process is the existence of (bilingual) parallel or comparable corpora. By parallel corpora we mean texts that are direct translations of each other, and by comparable corpora we mean texts with the approximate same content and referring to the same event. Such parallel/comparable texts contain a large amount of implicit information that can be used as a bridge to transfer linguistic knowledge and make it accessible for both languages. There has recently been a number of research projects with this focus (see e.g. (Yarowsky et al., 2001), (Hwa et al., 2002), (Riloff et al., 2002), (Alemu et al., 2004) and (Alemu et al., 2003)).

In this paper we present recent work with the focus of constructing a bilingual comparable corpus consisting of Amharic and English news texts. In doing so, we utilised a number of available software and data sources that were (mainly) found on the Internet. The Amharic and English news texts were collected from Walta Information Center, a private news agency located in Addis Ababa, Ethiopia, that makes

Alemu Argaw A. and Asker L. (2005).
WEB MINING FOR AN AMHARIC - ENGLISH BILINGUAL CORPUS.
In *Proceedings of the First International Conference on Web Information Systems and Technologies*, pages 239-246
DOI: 10.5220/0001228502390246

daily news in Amharic and English available through their web page[1]. Although this news agency provides Ethiopian news in both Amharic and English, only a small portion of the articles in the archive refer to the same event. Furthermore, the Amharic news texts are published using Ethiopic script and archived according to the Ethiopian calendar while the English news texts are archived according to the Gregorian calendar. Thus a major part of the work consisted of identifying the relevant comparable news items in the archive, transliterating the Amharic news items into an ASCII representation, and aligning them with their respective corresponding English version.

The work is motivated by the fact that 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. This problem, which is shared by a number of so called "low density languages" has proven to be a bottleneck when it comes to promote the use of computers and the Internet in local languages. It is difficult to develop new linguistic resources without access to already exsisting ones. A challenge has therefor been to show that it is possible to automatically align and create a bilingual comparable corpus even in the absence of these resources. It is our hope that by doing so, and by making available even a small parallel Amharic - English corpus, that it can provide a starting point for such transfer as well as provide a useful basis for a number of other higher level natural language processing activities for Amharic.

In order to align the Amharic and English news articles without the help of an electronic Amharic - English lexicon, we used the string closeness between words in the titles as a way to measure how similar two news items were. This was implemented using the Levenshtein Distance algorithm (also called edit distance) (Bendersky, 2004). In order to restrict the matching algorithm further, we only compared the titles of news items that were published on the corresponding same date and place. An experimental evaluation of the performance of the algorithm based on data for one year is presented below in Section 5.

2 BACKGROUND

2.1 Web Mining

The Internet has so far been predominated by English. Nevertheless, it shows great promise as a source for multilingual content due to the fact that texts in more and more languages are becoming available on the WWW, and the number is growing by the day. Resnik (Resnik, 1999) referred to numbers from the Babel survey of multilinguality on the Web[2] and presented estimated figures that as of June, 1997, there were on the order of 63,000 primarily non-English Web servers, ranging over 14 languages. A follow-up investigation of the amount of non-English web servers suggested that nearly a third contain information expressed in more than one language. Today, the total number of web servers has grown to more than 285 million (according to the Internet Systems Consortium's[3] Internet domain survey from July 2004, which is based on the number of hosts advertised in the DNS). Furtermore, the number of on-line non-English users are currently estimated to be 544.5 Million (latest figures are from 2004) compared to 295.4 Million English users (GlobalReach, 2004).

A number of researchers have developed web mining algorithms and tools to construct bilingual corpora from the web c.f. (Chen and Jian-Jun, 2000), (Ma and Liberman., 1999), (Yang and Li, 2002), (Resnik, 1998), (Resnik, 1999), (Resnik and Smith, 2003). The most common approaches in these systems have been to use structural resemblance, content analysis, or a combination of the two to find matching web pages. One example is the STRAND (Structural Translation Recognition for Acquiring Natural Data) system developed by Philip Resnik and colleagues (Resnik, 1998), (Resnik, 1999).

STRAND uses structural filtering to compare language pairs, linearizing the HTML structure of both documents and aligning the resulting sequences. STRAND's approach is to identify naturally occurring pairs of Web pages in parallel translation. STRAND locates pages that might be translations of one another via a number of different strategies, and filters out page pairs where the page structures diverge by too much. To attain this, it exploits an observation about the way Web page authors disseminate information in multiple languages: that when presenting the same content in two different languages, authors exhibit a very strong tendency to use the same document structure. Hence, STRAND is based on the insight that translated Web pages tend quite strongly to exhibit parallel structure, permitting their exploitation even without analyzing the content.

The original STRAND architecture used the AltaVista search engine to accomplished the first step by searching for two types of Web pages. For each web page that it finds, it looks for a parent page (one that contains hypertext links to different language versions of a document) and a sibling page (a page in one language that itself contains a link to a version of the same page in another language). When considering only parent and sibling pages, the identification of

[1] http://www.waltainfo.com

[2] http://www.isoc.org/

[3] http://www.isc.org

potentially translated pages is simply done by pairing the child or sibling pages. When all the pages on a site are under consideration, the matching is done by compairing URLs in order to exploit the fact that the directory structure on many Web sites have a parallel local directory structure for web pages that are translations of each other. Another possible feature that the authors consider for matching is the use of document lengths and the fact that texts which are translations of each other tend to be similar in length.

STRAND has been used to mine bilingual documents from the web for language pairs such as English-French, English-Spanish and the authors claim that rigorous evaluation using human judges suggests that the technique produces an extremely clean corpus - noise estimated between 0 and 8% even without human intervention (Resnik, 1999).

In (Resnik and Smith, 2003) STRAND is enhanced with content based similarity measures and applied to the Internet Archive[4] to obtain an English-Arabic parallel corpus of more than 1M tokens per language, with a precision of 0.95 and a recall of 0.99 over the extracted candidate pairs.

Due to the nature of the news data that we have been working with, our approach differs in a number of ways from the one used in STRAND and other similar tools for web mining of parallel corpora. Firstly, the structure and branching factor of the news archive does not allow for a straight forward structural matching. Secondly, the use of Ethiopic script and the lack of an electronic Amharic - English lexicon does not allow for a straight forward content based matching. Instead, we had to transliterate the Amharic texts and then use the following heuristics to guide the matching and alignment process:

1. Parallel texts are ususally published the same date
2. Parallel texts usually refer to the same place name
3. The titles of news items tend to have a high concentration of words that are relatively unique and specific for that particular news text. These words are usually nouns and a very high proportion of person and place names. These nouns are often similar across languages and can therefor be identified using fuzzy string matching.

2.2 Amharic

Amharic is the official government language spoken in Ethiopia. It is a Semitic Language of the Afro-Asiatic Language Group that is related to Hebrew, Arabic, and Syrian. Amharic, the syllabic language, uses a script which originated from the Ge'ez alphabet (the liturgical language of the Ethiopian Orthodox Church). The language has 33 basic characters with each having 7 forms for each consonant-vowel combination, and extra characters that are consonant-vowel-vowel combinations for some of the basic consonants and vowels. It also has a unique set of punctuation marks and digits. Unlike Arabic, Hebrew or Syrian, the language is written from left to right. Amharic alphabets are one of a kind and unique to Ethiopia.

According to the 1998 census (in Arthur Lynn's World Languages) Amharic is spoken widely through out different regions of Ethiopia: by over 17 million people as a first language and by over 5 million second language users. Some estimates indicate that Amharic is the mother-tongue of around 15 to 30 million Ethiopians. Manuscripts in Amharic are known from the 14th century and the language has been used as a general medium for literature, journalism, education, national business and cross-communication. A wide variety of literature including religious writings, fiction, poetry, plays, and magazines are available in the language.

Today a growing number of people use computers for various information processing purposes such as document writing and correction, storage and retrieval of Amharic texts and databases. Hence more and more documents (information) and databases in Amharic are becoming available in electronic form.

Amharic documents written in Ethiopic are available on the web, and the amount is increasing by the day. Ethiopic or Ethiopian script refers to the Ge'ez alphabet, and is the official writing system of Ethiopia. Different character encoding schemes and different keyboard layout are used to represent Ethiopic electronically. Some are unicode compliant (e.g. Visual Ge'ez, Ethiopia Jiret) while some are not. Although much effort has been made to have a standard for Ethiopic encoding, so far there is no standard and texts written in these different encoding schemes are not compatible with one another.

Although the amount of Amharic text on the web is growing in size, the availability of an equivalent translation of the texts in another language is still very limited.

2.3 The Ethiopian Calendar

The Ethiopian calendar runs approximately seven years and eight months behind the Gregorian calendar (the current year 2005 is 1997 in the Ethiopian calendar). The calendar is divided into 12 months of 30 days each and one 13th month consisting of five or six days depending on wether the current year is a leap year. The Ethiopian new year starts on September 11 (or September 12 on Gregorian leap years). For the purpose of finding corresponding dates in the Ethiopian and Gregorian calendars, a conversion table was extracted from a free trial version of the 7000 Years Calendar v1.4.1(JuneCalends, 2004). This ta-

[4] http://www.archive.org

ble was used to convert the dates from one calendar to the other and use this information to identify those articles that were written on the same date during the alignment process.

2.4 Parallel texts

Parallel and comparable corpora play an important role in machine translation and multilingual natural language processing. They represent resources for automatic lexical acquisition, they provide indispensable training data for statistical translation models, and they can provide the connection between vocabularies in cross-language information retrieval (Resnik and Smith, 2003). Recently, the trend to utilize parallel corpora to transfer linguistic resources from resource-rich languages such as English onto lesser supported languages has been referred to as "Cross-Language Projection" (Yarowsky et al., 2001). Comparable corpora can be used as resources for constructing parallel corpora, as well as resources for different corpus based linguistic, natural language processing and information retrieval experiments.

3 PREPROCESSING

3.1 The Data Set

There are very few parallel/comparable Amharic - English news texts available on the Internet. While there are a few sites (such as Ethiopian News Headlines[5]) that contain Ethiopian news in Amharic only, and others (such as Addis Tribune[6]) that contain Ethiopian news in English, we have only been able to find one, at the Walta information center[7] that contains a substantial amount of comparable versions of both Amharic and English News with URLs that give a clue for automatic alignement.

Not all texts at this site have a corresponding comparable version, but a portion of the news texts here describe the same event and are comparable (but not direct) translations of each other. These comparable news items are relatively few and hard to identify automatically but are in general archived under the corresponding dates. The Ethiopic news are archived using the Ethiopian calendar while the English news are using the Gregorian calendar, but most comparable news stories can be found within a couple of days from the corresponding date.

The archive at Walta contains Amharic news from the years 1993 until today (1997 Ethiopian calendar) while the English archive contains articles from the years 1999 until today). An estimation of the actual number of matching news items, based on manual inspection of 15 days of data, gives that approximately 10% of the news items refer to the same event. This would indicate that the archive in total contains around 900 matching news texts.

3.2 Downloading

We downloaded all English and Amharic news articles available at the Walta Information Center archives using a web crawler (an evaluation version of Offline Explorer Pro 3.5 from MetaProductsSoftware Corporation[8]). Since the pages at the archive contain a large number of links to other web pages that were irrelevant for our purposes, it was important to be able to filter and control exactly what was being retrieved.

The news tetxs at the Walta archive are structured in folders for Amharic news and English news with subfolders for each year, month and day respectively. The Amharic news is archived under folders according to the Amharic calendar while the English news is stored in folders with names according to the Gregorian calendar so a matching pair of news items would for example be stored under:

`EnNews/2002/feb/01Feb02/Feb1e8.htm`

and

`AmNews/1994/tir/24tir94/tir24a04.htm`

respectively. A particular day would typically contain between five and ten separate news items in each language. Since the file names do not contain sufficient information to identify the date (information about the year is missing) we downloaded all available English and Amharic articles from the archive while retaining the original folder structure.

3.3 HTML tag removal

When the file structure for the relevant news articles had been downloaded it was then flattened and the html code for each page was removed using a publicly available freeware (Emsa HTML Tag Remover v1.0 Build 20). This software allows for controlled removal of html tags as well as whitespace and other special characters from html files. An extra degree of control was required since the representation for the Amharic fonts includes some special characters (such as e.g. "{", "}" and "|") that would otherwise create problems for the transliteration if they had been removed. In addition to this it was important to preserve portions of the original file structure in order to simplify the parsing of the processed files into separate fields such as e.g. title, place name and date.

[5] http://www.ethiozena.net/

[6] http://www.addistribune.com

[7] http://www.waltainfo.com

[8] http://www.metaproducts.com

3.4 Transliteration

In order to simplify the analysis and matching of news texts and to have a unified representation of the Amharic texts, we decided to transliterate all Amharic texts into SERA (Yacob, 1996).

The Ethiopic script in the Amharic texts are represented using a variety of fonts. For the Amharic years 1993 until the first half of 1996, Visual Geez 2000 was the most common, while after that, a mixture of fonts have been used, which complicated the transliteration step.

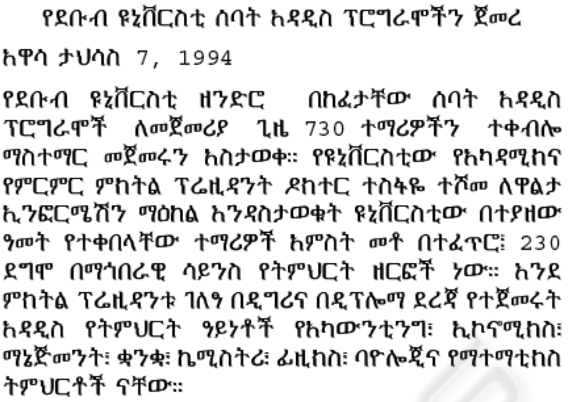
The transliteration was done using a file conversion utilty called g2 which is available in the LibEth package (LibEth is a library for Ethiopic text processing written in ANSI C[9]. g2 was made available to us by Daniel Yacob of the Ge'ez Frontier Foundation[10].

3.5 Restructuring

Most of the Amharic and English news texts have a semi-structured format that includes title, place name, date, newsagency, and body. In order to simplify the matching of news texts, we have preprocessed the news articles and stored them in an xml structure that identifies each of these fields separately. Figures 1 - 3 shows three different representations of the same news story, the original Amharic news text using Ethiopic script (Figure 1), the transliterated Amharic text (Figure 2), and the corresponding English news text (Figure 3).

4 ALIGNMENT

The news articles at Walta Information center are more comparable than parallel. Some are direct translations from Amharic to English or vice versa while others are news stories written by different reporters but describing the same event. The ones we tried to match in this experiment are all those that describe the same incident. The date the articles are written, the place the incident occured and the title of the articles are the major information sources used for the alignment. The basic assumption here is based on the fact that titles tend to be a highly summarised version of the news text and tend to have content words that are nouns or noun phrases which in turn are usually place names, person names, organization names, or dates, numbers etc. Hulth (Hulth, 2004) has 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. It seemed



የደቡብ ዩኒቨርስቲ ሰባት አዳዲስ ፕሮግራሞችን ጀመረ

አዋሳ ታህሳስ 7, 1994

የደቡብ ዩኒቨርስቲ ዘንድሮ በከፈታቸው ሰባት አዳዲስ ፕሮግራሞች ለመጀመሪያ ጊዜ 730 ተማሪዎችን ተቀብሎ ማስተማር መጀመሩን አስታወቀ። የዩኒቨርስቲው የአካዳሚክና የምርምር ምክትል ፕሬዚዳንት ዶክተር ተስፋዬ ተሾመ ለዋልታ ኢንፎርሜሽን ማዕከል እንዳስታወቁት ዩኒቨርስቲው በተያዘው ዓመት የተቀበላቸው ተማሪዎች አምስት መቶ በተፈጥሮ፤ 230 ደግሞ በማኅበራዊ ሳይንስ የትምህርት ዘርፎች ነው። እንደ ምክትል ፕሬዚዳንቱ ገለፃ በዲግሪና በዲፕሎማ ደረጃ የተጀመሩት አዳዲስ የትምህርት ዓይነቶች የአካውንቲንግ፣ ኢኮኖሚክስ፣ ማኔጅመንት፣ ቋንቋ፣ ኬሚስትሪ፣ ፊዚክስ፣ ባዮሎጂና የማተማቲክስ ትምህርቶች ናቸው።

Figure 1: Part of an Amharic news text represented in Ethiopic font

```
    yedebub yuniversti sebat adadis
           programocn jemere
awasa tahsas 7, 1994
yedebub  yuniversti  zendro  bekefetacew
sebat adadis programoc lemejemeriya gizE
730   temariwocn teqeblo    mastemar
mejemerun astaweqe:: yeyuniverstiw
yeakadamikna  yemrmr   mktl prEzidant
dokter tesfayE texome lewalta informExn
ma'Ikel Indastawequt yuniverstiw
beteyazew 'amet yeteqebelacew temariwoc
amst    meto betefeTro; 230   degmo
bema'heberawi sayns yetmhrt zerfoc new::
Inde mktl prEzidantu gele'Sa  bedigrina
bediploma dereja yetejemerut adadis
yetmhrt 'aynetoc yeakawnting, ikonomiks,
manEjment,  qWanqWa, kEmistri, fiziks,
bayolojina yematematiks tmhrtoc nacew::
```

Figure 2: The news text from Figure 1 (above) represented in SERA

```
  Debub University Introduces Seven New
           Field of Studies
Awassa December 16, 2001
The Debub University in Awassa disclosed
that it  had admitted  some 730 students
in  seven  new   field  of   studies  it
launched at degree and  diploma  levels
during   the   current   academic  year.
Academic and Research  Vice president of
the University, Dr. Tesfaye Teshome told
WIC  that  the  new  fields  of  studies
include       Accounting,      Economics,
Management,      Language,     Chemistry,
Physics, Biology and Mathematics.
```

Figure 3: Aligned English news text corresponding to Figure 1 (above)

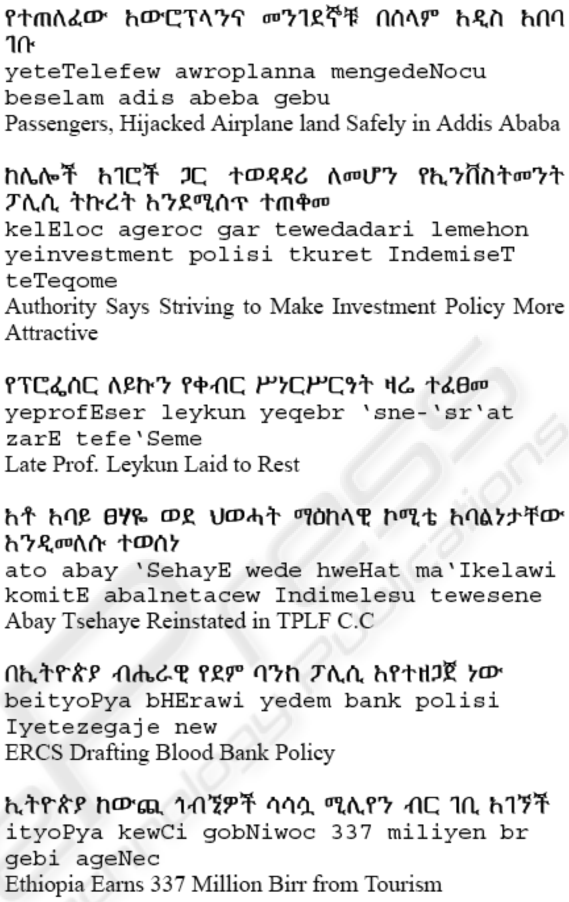
[9] http://libeth.sourceforge.net/

[10] http://www.ethiopic.org/

like a reasonable assumption to expect that news titles would contain a similar proportion of noun phrases. From the outset it would appear as if the Amharic and English titles are very different, since they are in different languages and using different alphabets. But once the Amharic version is transliterated, it becomes more apparent that many nouns and proper names are in fact quite similar (see Figure 4). This information can be used to match parallel Amharic English text without using any lexical resources that are hardly available for this language pair. Even if it was available, it would be very unlikely that the lexicon would contain an extensive list of proper names.

An algorithm was designed and implemented to automatically align documents that are comparable. The basic scheme was to retrieve the date information from the Amharic news articles, get the corresponding Gregorian date from the calendar conversion list, and try to align them with the English news articles from that same date. From the set of English articles with the same date, it gets the place information and matches those which have similar place names with that of the Amharic article under consideration. From the set of English articles with matching place names, it matches each word in the Amharic article's title with each word in the English article's title, and retrieves the one with the best edit distance score above a match threshold (specified by the user) as the best match. Words that are found to be matches in the English side are excluded from being matched with subsequent Amharic words. The number of words in the Amharic news title that have a match in the corresponding English title are then counted and divided by the total number of words in the longer title. The fuzzy string matching was done using a perl implementation of Levenshtein Distance(Bendersky, 2004).

Levenshtein distance (LD) is named after the Russian scientist Vladimir Levenshtein, who devised the algorithm in 1965. The metric is also sometimes called edit distance. Levenshtein distance is a measure of the similarity between two strings. The distance is the number of deletions, insertions, or substitutions required to transform a sourse string into a target string. For example if the source (s) is "addis" and the target(t) is "addis", then LD(s,t) = 0, because no transformations are needed. The strings are already identical. If s is "adis" and t is "addis", then LD(s,t) = 1, because one insertion is sufficient to transform s into t. The greater the Levenshtein distance, the more different the strings are. In our implementation, to calculate the edit distance score, we divided the Levenshtein distance for a word, by the length of the same word, in order to normalise and compensate for the fact that longer words tend to have a larger edit distance.



የተጠለፈው አውሮፕላንና መንገደኞቹ በሰላም አዲስ አበባ ገቡ
yeteTelefew awroplanna mengedeNocu beselam adis abeba gebu
Passengers, Hijacked Airplane land Safely in Addis Ababa

ከሌሎች አገሮች ጋር ተወዳዳሪ ለመሆን የኢንቨስትመንት ፖሊሲ ትኩረት አንደሚሰጥ ተጠቆመ
kelEloc ageroc gar tewedadari lemehon yeinvestment polisi tkuret IndemiseT teTeqome
Authority Says Striving to Make Investment Policy More Attractive

የፕሮፌሰር ለይኩን የቀብር ሥነርሥርዓት ዛሬ ተፈፀመ
yeprofEser leykun yeqebr 'sne-'sr'at zarE tefe'Seme
Late Prof. Leykun Laid to Rest

አቶ አባይ ፀሃዬ ወደ ህወሓት ማዕከላዊ ኮሚቴ አባልነታቸው አንዲመለሱ ተወሰነ
ato abay 'SehayE wede hweHat ma'Ikelawi komitE abalnetacew Indimelesu tewesene
Abay Tsehaye Reinstated in TPLF C.C

በኢትዮጵያ ብሔራዊ የደም ባንክ ፖሊሲ አየተዘጋጀ ነው
beityoPya bHErawi yedem bank polisi Iyetezegaje new
ERCS Drafting Blood Bank Policy

ኢትዮጵያ ከውጪ ጎብኚዎች ሳሳኋ ሚሊየን ብር ገቢ አገኘች
ityoPya kewCi gobNiwoc 337 miliyen br gebi ageNec
Ethiopia Earns 337 Million Birr from Tourism

Figure 4: Examples of some matching news items

5 EXPERIMENTAL EVALUATION

In order to test how efficient the fuzzy string matching would be for aligning news articles, we conducted a set of experiments where we used a subset of the data, from the (Gregorian) year 2001. The alignment process is based on two assumptions. Firstly, that a subset of words with the same meaning in the two languages will also have a similar (or identical) spelling. This is especially common for names of places and people, but also for loan words and cognates with the same etymological origin. Secondly, we assume that this subset of similar words also are words with high information value that tend to occur in the titles of new stories.

For the experiments, we have therefore investigated how similar words with the same meaning and origin tend to be. This was done by varying the word match threshold. The word match threshold is a value between 0 and 1 that represents the average number of edit distance operations per character that are required to transform one word into another.

Table 1: Results for the 2001 data set

Title threshold	Word threshold	Matches	Correct	Precision
0.15	0.4	153	89	0.58
0.20	0.4	44	33	0.75
0.25	0.4	13	12	0.92
0.30	0.4	5	5	1.00
0.35	0.4	2	2	1.00
0.15	0.5	225	127	0.56
0.20	0.5	77	57	0.74
0.25	0.5	30	29	0.97
0.30	0.5	13	13	1.00
0.35	0.5	3	3	1.00
0.20	0.6	302	129	0.43
0.25	0.6	122	74	0.61
0.30	0.6	62	50	0.81
0.35	0.6	38	35	0.92

We also investigated how large the portion of matching words in the titles would be for two news items that describe the same event. This was done by varying the title match threshold. The title match threshold is a value between 0 and 1 that represents the proportion of title words that would match in two news items that describe the same event.

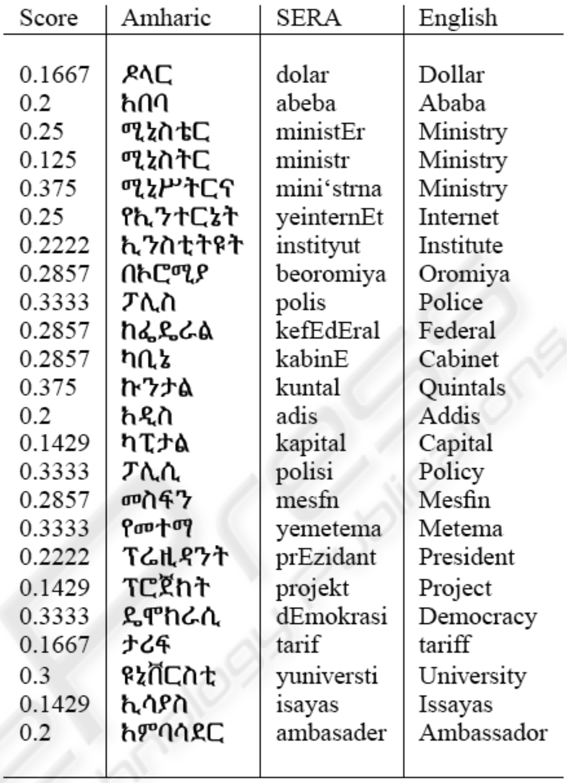
A title match threshold of 0.25 would for example mean that at least 25% of the words in the title would have an edit distance match, and a word match threshold of 0.5 would mean that not more than 50% of the characters of a word are allowed to be changed when matching two words.

The 2001 experiments were conducted on a data set consisting of 1923 English news articles published during the year 2001 (according to the Gregorian calendar), and 1219 Amharic articles published during the time interval between the 7th month of 1993 and the 4th month 1994 (Ethiopian calendar). The Amharic articles corresponding to 2001 should have started from the 5th month of 1993, but what is available in the archive starts from the 7th month of 1993.

The result of the matching was manually evaluated in order to calculate the precision values. Recall has not been calculated (due to the amount of manual work that this would require) but we have done an estimate for a randomly selected 15 days of data (February 1 - 15, 2002). The amount of matching news articles in this subset is 10 out of 98, or approximately 10%. In conducting the experiments, we aimed at finding corresponding comparable English articles for at least 10% of the total amount of Amharic articles. Examples of some news items with matching titles are shown in Figure 4.

As can be seen from the results reported in Table 1, the more constrains there are, the better the precision of the matches is, at a cost of very limited recall. When the word threshold was set to 0.5 or 0.4 the precision was 100% for title threshold values of 0.3 and 0.35. These same title threshold values give a lesser precision and better recall with a less constrained word threshold value of 0.6. With a more constrained word threshold value of 0.4, all experiments show an increase in precision and decrease in recall compared to the experiments with word threshold values of 0.5 and 0.6.

Table 2: Some Amharic and English words and their edit distance score



Score	Amharic	SERA	English
0.1667	ዶላር	dolar	Dollar
0.2	አበባ	abeba	Ababa
0.25	ሚኒስቴር	ministEr	Ministry
0.125	ሚኒስትር	ministr	Ministry
0.375	ሚኒሥትርና	mini‘strna	Ministry
0.25	የኢንተርኔት	yeinternEt	Internet
0.2222	ኢንስቲትዩት	instityut	Institute
0.2857	በኦሮሚያ	beoromiya	Oromiya
0.3333	ፖሊስ	polis	Police
0.2857	ከፌዴራል	kefEdEral	Federal
0.2857	ካቢኔ	kabinE	Cabinet
0.375	ኩንታል	kuntal	Quintals
0.2	አዲስ	adis	Addis
0.1429	ካፒታል	kapital	Capital
0.3333	ፖሊሲ	polisi	Policy
0.2857	መስፍን	mesfn	Mesfin
0.3333	የመተማ	yemetema	Metema
0.2222	ፕሬዚዳንት	prEzidant	President
0.1429	ፕሮጀክት	projekt	Project
0.3333	ዴሞክራሲ	dEmokrasi	Democracy
0.1667	ታሪፍ	tarif	tariff
0.3	ዩኒቨርስቲ	yuniversti	University
0.1429	ኢሳያስ	isayas	Issayas
0.2	አምባሳደር	ambasader	Ambassador

While conducting the alignment experiments, the words that are returned as closest matches could be used in further alignment experiments. Some examples of correctly paired words are given in Table 2.

6 CONCLUSIONS

We have shown how fuzzy string matching between Amharic and English words can be a sufficient and useful way to align texts in the two languages. Although the texts at first sight appear to be completely different, it is possible to identify a large portion of words in transliterated texts that are similar enough to allow for a fuzzy string matching approach to aligning texts. We used edit distance as a way to measure how similar two strings were and calculated a score that would also take word length into consideration.

Since the Amharic and English news texts in this study are comparable rather than parallel (direct translations of one another), the algorithm did not use document length as a feature in the alignment process. We believe that under these circumstances, the document length could be more confusing than helpful for the alignment process. When aligning potentially parallel data however, the length could be an important feature.

The content of the body of the text (the news article) could of course also be used in the alignment process instead of using the title only. The body of the text has often many occurences of names and numbers that would help the alignment. When analyzing the text body, resources such as lexica (e.g. for word by word translation), stop word list (to remove non content bearing words that appear in many of the articles), morphological analysis or stemming (to consider the root word only) etc. may be used.

Machine learning could also be used to improve the fuzzy matching by finding more likely character substitutions from known matching word pairs, and assigning them a different weight when calculating the word matching score. It would also be possible to incorporate an improved number and date conversion that would allow several different formats (digit or text representation) for these items.

Under all circumstances, when used as a semi automatic tool, the existing algorithm gives an acceptable performance in relation to the amount of work that would otherwise be required to align the news items manually. As an example, the parameter settings 0.15 and 0.5 finds 225 suggested matches out of which 127 are correctly aligned news items (see Table 1 above). The 127 correctly aligned texts are resonably close to the estimated total number of possible matches (10% of the 1219 Amharic articles in the test set), and at the same time the amount of manual work required to identify the 98 incorrectly aligned news pairs in this group is substancially less than what would have been required to do it completely from scratch without the help of the system.

It is our hope that by demonstrating the feasibility of our approach, we will inspire additional work on creating parallel corpora and other linguistic resources for Amharic as well as for many more of the worlds "low density languages".

REFERENCES

Alemu, A., Asker, L., and Eriksson, G. (2003). An empirical approach to building an amharic treebank. In *Proceedings of TLT-2003*.

Alemu, A., Asker, L., and Eriksson, G. (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-2004*.

Bendersky, E. (2004). Levenshtein distance algorithm: Perl implementation. http://www.merriampark.com/ldperl.htm, accessed Jan 31, 2004.

Chen, J. and Jian-Jun, N. (2000). Automatic construction of parallel english-chinese corpus for cross-language information retrieval. In *Proceedings of the Sixth Conference on Applied Natural Language Processing*.

GlobalReach (2004). Global internet statictics (by language). http://global-reach.biz/globstats/index.php3.

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

Hwa, R., Resnik, P., Weinberg, A., and Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of ACL-02*.

JuneCalends (2004). 7000 years calendar v1.4.1. http://www.junecalends.com/7000.html, accessed Jan 31, 2004.

Ma, X. and Liberman., M. (1999). Bits: A method for bilingual text search over the web. In *Proceedings of Machine Translation Summit VII*.

Resnik, P. (1998). Parallel strands: A preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98*.

Resnik, P. (1999). Mining the web for bilingual text. In *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3).

Riloff, E., Schafer, C., and Yarowsky, D. (2002). Inducing information extraction systems for new languages via cross-language projection. In *Proceedings of COLING-02*.

Yacob, D. (1996). System for ethiopic representation in ascii (sera). http://www.abyssiniacybergateway.net/fidel/.

Yang, C. C. and Li, K. W. (2002). Mining english/chinese parallel documents from the world wide web. In *Proceedings of the 11th International World Wide Web Conference*.

Yarowsky, D., Ngai, G., and Wicentowski., R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT-01*.