Oromo-English Information Retrieval Experiments at CLEF 2006

Kula Kekeba Tune and Vasudeva Varma
Language Technologies Research Center
International Institute of Information Technology, Hyderabad, India
(kuulaa@gmail.com, vv@iiit.ac.in)

Abstract

Our Cross-Language Information Retrieval research group, part of Search and Information Extraction Lab (SIEL) at Language Technologies Research Center (LTRC) of IIIT-Hyderabad has participated this year, for the first time, in CLEF campaign. We took part in ad hoc track by conducting various bilingual information retrieval experiments for three different languages: Oromo-English, Hindi-English, and Telugu-English. In this paper we describe our Oromo–English information retrieval experiments at CLEF'06. The main objective of all Oromo-English retrieval experiments was to assess the over all performance of our dictionary-based CLIR system by using different fields of Afaan Oromo topics. We submitted three different official runs, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN). Since Afaan Oromo is one of the linguistic resource scarce languages, very limited linguistic resources such as Oromo-English dictionary and Afaan Oromo stemmer have been used in our experiments. Yet we found the results that we have achieved at CLEF'06 for all of our Oromo-English experiments very encouraging.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-Language Retrieval, Afaan Oromo, Oromo, Bilingual Information Retrieval

1. Introduction

This year we had made our debut participation in CLEF ad hoc track from Language Technologies Research Center (LTRC) of IIIT-Hyderabad, India. Our main purpose was to obtain hands-on experience in joint evaluations of cross-language information retrieval by conducting experiments in three different languages, i.e. Oromo-English, Hindi-English and Telugu-English. This paper presents a brief description of our Oromo-English CLIR experiments at CLEF 2006.

Oromo (also known as Afaan Oromo) is one of the major Languages that are widely used in Ethiopia [3]. Currently it is an official language of Oromia (which is the largest region/state in Ethiopia). Unlike Amharic, (another major language and official language of Ethiopia) which belongs to Semitic family languages, Afaan Oromo is part of the Lowland East Cushitic group within the Cushitic family of the Afro-Asiatic phylum [1, 3]. In this Cushitic branch of the Afro-asiatic language family Afaan Oromo is considered as one of the most extensive languages among the forty or so Cushitic languages [2]. It is a common mother tongue for Oromos, who are the largest ethnic group in Ethiopia, at 32.1% of the population according to the 1994 census. It is spoken by approximately 24-25 million Oromos [4] within Ethiopia.

With regard to the writing system, Qubee (Latin-based alphabet) has adopted and become the official script of Afaan Oromo since 1991. Currently, Afaan Oromo is widely used as both written and spoken languages in Ethiopia and some neighboring countries, including Kenya and Somalia. Besides being an official language of Oromia, Afaan Oromo is the instructional medium for primary and junior secondary schools throughout the region and its administrative zones.

Like a number of other African and Ethiopian languages, Afaan Oromo has a very rich morphology. It has the basic features of *agglutinative* languages where all bound forms (morphemes) are affixes. Major word formations in Afaan Oromo involve affixation, reduplication and compounding [5]. For instance, both Oromo nouns and adjectives are inflected for number, gender and person while verbs are also inflected for gender, person, number and tenses. Moreover, possessions, cases and article markers are often indicated through affixes. Obviously, these high inflectional forms and extensive derivational features of the language are presenting various challenges for text processing and information retrieval experiments in Afaan Oromo. Our major aim in participating in CLEF'06 is to assess the level of performance that can be achieve by experimenting with simple dictionary-based CLIR system, like Oromo-English, where sophisticated language resources and IR techniques are not employed.

The rest of this paper is organized as follows: Section 2 describes our experimental setup including basic procedures and approaches that have been used in our official runs. Section 3 presents the results obtained for our three official runs while section 4 summarizes our conclusions.

2. Experimental Setup

As indicated in the above section, we have participated in ad hoc track of CLEF campaign by submitting three official runs in Oromo-English bilingual task. We used the English test collection (provided by CLEF) from two newspapers, the *Glasgow Herald* and the *Los Angeles Times*. We have used a dictionary-based CLIR method that is similar to some of bilingual task experiments at the previous CLEF campaigns [6, 7, 8]. Oromo-English dictionary was adopted and developed from hard copies of human readable bilingual dictionaries by using OCR technology. This resource is used to translate Oromo topics into a bag of words English queries. Then these translated English queries are submitted to text retrieval engine. Accordingly, we had conducted various experiments and submitted three official runs, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN). Lucene [8], an open source text search engine that is mainly based on *vector space model* [9] was adopted and used for indexing and retrieval of the documents.

2.1 Stop word Lists and Stemming

In order to define Oromo stop words, we first created a list of the top 350 most frequent words found in 1.2 million words Afaan Oromo text corpus. Then we added pronouns, conjunctions, prepositions and other similar functional words in Afaan Oromo. Accordingly, we have obtained and used about 580 stop words of Afaan Oromo in conducting our experiments. Once these high-frequency words were removed from Oromo topics, we applied a light stemming algorithm in order to conflate word variants into the same stem or root. Similarly, after query topics are translated into English both topics and documents were also stopped through Lucene search engine.

A number of previous works on CLIR have indicated that languages that are morphologically rich can benefit from morphological analysis such as stemming and lemmatization [10]. Since Afaan Oromo is one of such morphologically rich languages and stemming is often language dependent, we have developed a rule based suffix-stripping algorithms focusing on very common inflectional suffixes. This light stemmer is designed to automatically remove frequent inflectional suffixes attached to base-words of Afaan Oromo. Some of the common suffixes that have been considered in our current light stemmer include gender (masculine, feminine), number (singular or plural), case (nominative, dative), possession morphemes and other related morphological features in Afaan Oromo.

2.2. Query Translation

Initially, the original CLEF query set for English topics were manually translated into Oromo topics by a group of translators who are native speakers of Afaan Oromo. We then translated the resulting Oromo topics back into English queries using Oromo-English dictionary that was adopted and developed from human readable bilingual dictionaries by using OCR technology. Currently this bilingual dictionary is consists of about 12,800 Oromo entries including base words and their derivational variants.

After eliminating of stop words, the stemmed keywords words of Oromo topics were automatically looked up for all possible translations in the bilingual dictionary. Therefore, the resulting English query is a simple bag of words, taking into account all the possible translation of keywords found in the bilingual dictionary. One of the major problems in this translation process is unknown or unmatched words which are about 120. While most of these words are proper names, few of them are foreign words. Since the transliteration of some of these unmatched proper names or foreign words have nearly identical spelling with the corresponding English words (e.g. Iraaq, Kurdi, Buush, fiilmi) they words are directly copied and added to the English query for partial matching by the search engine. The transliteration of the rest (about 68) of the proper names which are more complex were manually modified and added in a separate dictionary. Handling of phrasal and compound words are not considered in our current experiments.

3. Experimental Results

In this section we describe the results of our experiments. We had conducted and submitted three official runs, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN) for Oromo-English bilingual task in the ad-hoc track of CLEF'06. The Mean Average Precision (MAP) and Geometric Average Precision (GAP) scores for our three runs are shown in Table 1. The total number of relevant documents (Relevant-tot.), the retrieved relevant documents (Rel.Ret.), and the non-interpolated average precision (R-Precision) are also summarized in Table 1. Table 2 shows summary of Recall-Precision results for the three runs.

Run-Label	Relevant-tot.	Rel. Ret.	MAP	R-Prec.	GAP
OMT	1,258	870	22.00%	24.33%	7.50%
OMTD	1,258	848	25.04%	26.24%	9.85%
OMTDN	1,258	892	24.50%	25.72%	9.82%

Table 1. Summary of average results for the three runs

Recall	OMT	OMTD	OMTDN
0%	48.73%	58.01%	59.50%
10%	39.93%	47.75%	46.45%
20%	34.94%	42.33%	37.77%
30%	30.05%	32.15%	31.17%
40%	26.41%	28.55%	28.27%
50%	22.98%	24.90%	24.72%
60%	18.27%	20.19%	19.40%
70%	15.10%	16.59%	15.61%
80%	11.76%	12.87%	12.70%
90%	8.58%	8.37%	8.56%
100%	6.56%	6.05%	6.58%

Table 2. Recall–Precision scores for the three runs

Table 1 reveals OMTD (title and description) run and OMTDN (title, description and narration) run have achieved almost the same level of performance (with about MAP of 25 %). The title run has slightly lower performance with MAP of 22%. We feel this is due to the fact that most of the title fields in Afaan Oromo topics were very short. It is also worth noting that the geometric average precision (GAP) score of title run is also slightly below the GAP score of the other two runs.

5 Conclusions and Future Works

Based on the results of our Oromo-English experiments, we attempted to show how very limited language resources can be used in a bilingual information retrieval setting. Since this is the first time we participated in CLEF campaign, we concentrated on evaluation of the over all performance of the Oromo-English CLIR system which is being developed at our research center. We feel we have obtained reasonable and significant average results for all of the three official runs, given the limited CLIR resources we have used in our experiments. However, we know the fact that there is a lot of room for improvement of the performance our Oromo-English CLIR system. Currently we are working on evaluation of different components of the CLIR system to identify the effects of stop words and light stemmer of Afaan Oromo. Automatic query expansion by using pseudo-relevance feedback, proper names handling and application some disambiguation methods are some the tasks that we will consider in our future experiments.

REFERENCES

1. Baye Yimam. The Phrase Structure of Ethiopian Oromo. Ph.D. Thesis. School of Oriental and African Studies, University of London, 1986.

2. H.A. Stroomer. Comparative Study of Southern Oromo Dialects in Kenya: Phonology, Morphology and Vocabulary. Hamburg: Burke, 1987.

3. Abara Nefa. Long Vowels in Afaan Oromo: A Generative Approach. M.A. Thesis. School of Graduate Studies, Addis Ababa University, 1988.

4. Oromo Language. http://en.wikipedia.org/wiki/Oromo_language

5. Gumii Qormaata Afaan Oromoo, Komishinii Aadaaf Turizmii Oromiyaa. Caasluga Afaan Oromoo, Jildii – 1. Finfinnee, 1995 (E.C.).

6. Atelach Alemu Argaw, et. al. Dictionary Based Amharic – English Information Retrieval, 2004. http://www.clef-campaign.org/2004/working_notes/WorkingNotes2004/13.pdf

7. _____. Dictionary - Based Amharic – French Information Reyrieval, 2005. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/argaw05.pdf

8. Marin Carpuat and Pascale Fung. CLEF 2001 Bilingual Task: Simple Dictionary-Based Query Translation, 2001. http://www.ercim.org/publication/ws-proceedings/CLEF2/carpuat.pdf.

9. Apache Lucene: http://lucene.apache.org.

10. Jacques Savoy and Pierre-Yves Berger. Report on CLEF-2005 Evaluation Campaign: Monolingual, Bilingual, and GIRT Information Retrieval, 2005. http://www.clef-campaign.org/2005/working_notes/workingnotes2005/savoy05.pdf