# Oromo-English Information Retrieval Experiments at CLEF 2007

Kula Kekeba Tune and Vasudeva Varma
Language Technologies Research Centre
IIIT-Hyderabad, India.
kuulaa@gmail.com, vv@iiit.ac.in

**Abstract**

In this paper we describe our Oromo-English retrieval experiments that we have conducted at IIIT-Hyderabad (India) and submitted to the ad hoc retrieval task of CLEF 2007. We participated in the bilingual subtask of CLEF campaign for the second time by designing and submitting four official runs. The experiments differ from one another in terms of topic fields used for query construction and the application of stemmer for normalization of query terms. One of our major objectives was to assess the overall performance of our dictionary-based Oromo-English CLIR system on a new English test collection that has been provided by CLEF this year. We are also interested in exploring and assessing the impacts of Afaan Oromo light stemmer on the overall performances of our experimental CLIR system. After a brief description of the research contexts of our Oromo-English CLIR system, we will present and discuss the evaluation results of our official runs.

**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; H.3.4 Systems and Software; H.3.7 Digital Libraries

**General Terms**

Languages, Measurement, Performance, Experimentation

**Keywords**

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

## 1 Introduction

In this paper we present a report on our Oromo-English retrieval experiments that we had conducted and submitted to the ad hoc track of CLEF 2007. In our second participation in the bilingual task of CLEF this year, we had designed and submitted four CLIR experiments using Afaan Oromo as source (query) language for retrieval of relevant documents from a large size of English test collection. The experiments differ from one another in terms of topic fields that are used for query construction and the application of Afaan Oromo stemmer for normalization of Oromo query terms. Due to lack of language processing resources and information retrieval tools that are appropriate for Afaan Oromo, only limited linguistic resources such as Oromo-English dictionary, Oromo light stemmer and stopwords [3] that have been designed and developed at our research center were used in conducting the experiments. Basically, we are motivated by the needs and challenges of designing and developing an experimental CLIR system for Afaan Oromo not only because it is one of the major African languages but because it is also one of the less resourced and indigenous languages of Africa. In our current Oromo-English CLIR study we have mainly focused on investigating and assessing the performance levels that we could achieve by designing and employing the scarcely available language resources of Afaan Oromo.

Since one of the driving forces behind our participation in CLEF 2007 has been to explore the effects of Afaan Oromo light stemmer on the performances of our CLIR system, we have designed and submitted the experiments in two sets. While one of our experiments was conducted and submitted without employing Afaan

Oromo light stemmer, the other three experiments were carried out and submitted with the application of the light stemmer. Moreover, all Oromo topic fields were used for query construction in the experiment that had been submitted without employing the light stemmer, (i.e. NOST-OMTDN07). We used our existing CLIR platform that had been reported in our previous works [3, 6] in conducting the experiments. In the subsequent sections we will briefly describe the major procedures that we have adopted in designing and conducting our Oromo-English CLIR experiments together with the evaluation results of the official runs.

The rest of this paper is organized as follows. Section 2 presents an overview of the linguistic features of Afaan Oromo from the point of view of CLIR application. Section 3 provides a brief description of Afaan Oromo light stemmer. Section 4 describes our experimental setup while section 5 summarizes and discusses the evaluation results that we have obtained for our official runs. Finally, section 6 provides our general concluding remarks.

## 2 Afaan Oromo and Its Morphology

Oromo (also often referred to as Afaan Oromo) is one of the major African languages that is widely spoken and used in most parts of Ethiopia and some parts of other neighbor countries in the horn of Africa. Currently, it is an official language of Oromia state (which is the largest Regional State among the current Federal States in Ethiopia). Afaan Oromo belongs to the Lowland East Cushitic group in the Cushitic family of the Afro-Asiatic phylum [1, 2]. It is the most prominent Cushitic family language that is closely related to Somali and Sidama [7]. Although it is difficult to identify the actual number of Afaan Oromo speakers (as a mother tongue) due to lack appropriate current information sources, according to some earlier general information sources it is estimated that Afaan Oromo is spoken by more than 25 million Oromos within Ethiopia. With regard to the writing system, Qubee (a Latin-based alphabets) has been adopted and become the official script of Afaan Oromo since 1991. Currently, Afaan Oromo is widely used as both written and spoken language in Ethiopia and some neighboring countries including Kenya and Somalia.

Like a number of other African and Ethiopian languages, Afaan Oromo has a very complex and rich morphology. It has the basic features of agglutinative languages involving very extensive inflectional and derivational morphological processes. In agglutinative languages like Afaan Oromo, most of the grammatical information is conveyed through affixes, (i.e. prefixes and suffixes) attached to the root or stem of words. Although Afaan Oromo words have some prefixes and infixes, in this paper we will focus on Oromo suffixes since they are the predominant morphological features in the language. Almost all Oromo nouns in a given text have person, number, gender and possession markers which are concatenated and affixed to a stem or singular noun form. In addition, Afaan Oromo noun plural markers/forms can have several alternatives. For instance, in comparison to the English noun plural marker *s(-es)*, there are more than ten major and very common plural markers in Afaan Oromo including: *-oota, -ooli, -wwan, -lee, -an, een, -eeyyii, -oo,* etc.). As an example, the Afaan Oromo singular noun "*mana*" (house) can take the following different plural forms: *manoota (mana + aota), manneen (mana + een), manawwan (mana + wwan).* In certain more complicated situations Oromo noun may take more than one plural markers concatenating and suffixing them one after the other, just to indicate the plural form of the noun as in: *manneenota (mana + een + ota)* or *manneenotaawwan (mana + een + ota + wwan).* The construction and usages of such alternative affixes and attachments are governed by the morphological and syntactic rules of the language. Oromo nouns have also a number of different cases and gender suffixes depending on the grammatical level and classification system used to analyze them.

Few examples of frequent gender markers in Afaan Oromo include: *-eessa/-eetii, -a/-ttii* or *-tu*. For instance, singular noun *obboleessa,* i.e. *obbol + eessa* (*M,* brother) vs. singular noun *obboleettii*, i.e. *obbol + eettii* (*F,* sister) and singular noun *garba,* i.e. *garb + a* (*M,* servant) vs. female singular noun *garbitti,* i.e. *garb + itti* (*F,* servant). And the plural noun *garboota,* i.e. *garb + a + oota (M,* servants) vs. plural noun *garbtoota,* i.e. *garb + iti + oota (F,* servants*)*. Like wise, Afaan Oromo adjectives have cases, person, number, gender and possession markers similar to Oromo nouns. Afaan Oromo verbs are also highly inflected for gender, person, number, tenses, voice and transitivity. Furthermore, prepositions, postpositions and article markers are often indicated through affixes in Afaan Oromo. Since Afaan Oromo is morphologically very productive, derivation, reduplication and compounding are also common in the language [4]. Obviously, these extensive inflectional and derivational features of the language are presenting various challenges for text processing and information retrieval tasks in Afaan Oromo. In information retrieval, the abundance of different word forms and lexical variability may result in a greater likelihood of mismatch between the forms of a keyword in a query and its variant forms found in the document index database(s). In the context of CLIR this may leads to a serious

mismatch problem between query terms and citation forms of vocabulary entries found in the bilingual dictionaries that are commonly used for cross language information retrieval.

## 3 Overview of Afaan Oromo Stemmer

Applications of certain level of morphological (linguistic) analysis and natural language processing tools are often assumed to be very essential in CLIR experiments of morphologically rich and agglutinative languages like Afaan Oromo. A number of previous research works, including [5, 10] have indicated the fact that CLIR applications in morphologically rich languages can benefit from stemming and lemmatization of query terms. As mentioned in the foregoing section of this paper, since Afaan Oromo is one of the morphologically rich languages and the process of stemming is often language dependent, we have designed and developed a rule-based light stemmer for Afaan Oromo focusing on its major inflectional and attached affixes. Since we are using a bilingual dictionary for query translation in our Oromo-English CLIR system, the dictionary lookup process requires that the Afaan Oromo query terms should be first stemmed and represented by their normalized and citation forms.

Broadly, it is possible to categorize the major types of suffixes in Afaan Oromo into three basic groups: *derivational*, *inflectional*, and *attached* suffixes [7]. Afaan Oromo *attached suffixes* are particles or postpositions like *-arra, -bira, -irra, -itti* and *-dha* while *inflectional suffixes* comprises the most frequent and dominant suffixes such as *–n, -lee, -een, -icha, -tu, -oo, -oota* and *-wwan*. Oromo *derivational suffixes* such as *-achuu, -eenyaa, -ina* and *-ummaa* are often used for formation of a new words in the language following the stems or base forms of Oromo words. Based on our current linguistic analysis and observations of Afaan Oromo syntax and morphological features, the most common order/sequence of the above major three Afaan Oromo suffixes (within a given word) is: *<stem><derivational suffixes><inflectional suffixes><attached suffixes>*. Thus, our Afaan Oromo stemmer is expected to remove (from the right end of a given word) first all the possible *attached suffixes*, then *inflectional suffixes* and finally *derivational suffixes* step by step. To facilitate this task, we have identified and constructed three different suffix clusters with respect to the above three major types of suffixes in Afaan Oromo.

Our current rule-based light stemmer is mainly designed to remove the most frequent attached and inflectional suffixes of Afaan Oromo from a given word (query term). Some of the most common suffixes that have been considered and handled by this light stemmer include gender (masculine, feminine), number (singular or plural), cases (nominative, dative), possession and other related bound morphemes of Afaan Oromo words. In addition, we have also used a stopword lists that we have created by using Oromo text corpus to facilitate the efficiency of our stemming algorithm and CLIR system. More detailed descriptions of these procedures were given in [3].

## 4 Experimental Setup

### 4.1 Query Processing and Translation

As indicated earlier, our dictionary-based Oromo-English CLIR system is based on query translation techniques. Initially, the original CLEF topic sets of English were manually translated into Oromo topic sets by a group of translators who are native speakers of Afaan Oromo. We then automatically translated these Oromo topic sets back into English queries using Oromo-English dictionary that was adopted and developed from human readable (printed) bilingual dictionaries. After tokenization, stopword elimination and stemming of Oromo topics (through the procedures we have described in the foregoing section), the stemmed keywords of Afaan Oromo query terms were automatically looked up in Oromo-English bilingual dictionary to identify all possible translations. In other words, since our current medium size bilingual dictionary has limited number of definitions for most of its vocabulary entries, we used all translated senses of Oromo query terms that are found in the dictionary. Therefore, the resulting translated English queries could be a set of terms (with multiple senses), which might have alternative or complementary English meanings that can serve as one means of query expansion.

However, some of the Afaan Oromo query terms may not be found in the bilingual dictionary since these term are either proper names or words borrowed from foreign languages or valid Afaan Oromo words which did not just occur in the dictionary. In some of the cases, the dictionary lookup for a given term might fail because of improper stemming or suffix removal. We have designed and used a set of heuristic rules for modification and translations of more complex and difficult Oromo query terms. Finally, the rest out-of-dictionary terms were selected and handled through automatic fuzzy matching and edit distance approaches that have been used in many CLIR research works including [8].

## 4.2 Retrieval Setup

We have adopted and used Apache Lucene [9], an open source text search engine for indexing and retrieval of the target test collections, i.e. English documents. Since Lucene is designed based on a vector space model, our document ranking is achieved through TF-IDF ranking algorithm that is based on a standard vector space model. We had designed and conducted four different retrieval experiments using Afaan Oromo as source (query) language for retrieval of relevant documents from a large size English text collection. Our experiments differ from one another in terms of topic fields that are used for query construction and the application of Afaan Oromo stemmer for normalization of Oromo query terms. Since we are interested in exploring and assessing the impacts of Afaan Oromo light stemmer on the overall performances of our CLIR system, we have designed and submitted our experiments in two sets. One experiment (i.e. NOST-OMTDN07) was conducted without employing Afaan Oromo light stemmer to serve as a baseline against the other three official runs. The rest three experiments (official runs) were conducted with the application of our light stemmer. Table 1 provides summary of our four official runs.

| Run-Id | Used Topic Fields | Stemming | Run Description |
|---|---|---|---|
| OMT07 | Title | Yes | Title Query Run |
| OMTD07 | Title and Description | Yes | Title and Description Query Run |
| OMTDN07 | Title, Description and Narrative | Yes | Title, Description and Narrative Query Run |
| NOST_OMTDN07 | Title, Description and Narrative | No | Title, Description and Narrative Query Run without Stemming |

Table 1. Summarized descriptions of the four official runs

## 5 Evaluation Results and Discussions

In this section we will present and discuss the evaluation results of our official runs that we have obtained from CLEF 2007. Table 2 shows the performances of our three different stemmed Oromo queries in terms of mean average precision (MAP) and R-Precision (R-Prec) scores. Average Precision scores after retrieval of the top 10 and 20 documents (i.e. P@10 and P@20) are also presented in the table.

| Run-Id | MAP (%) | R-Prec. (%) | P@10 (%) | P@20 (%) |
|---|---|---|---|---|
| OMT07 | 24.20 | 26.24 | 33.80 | 28.80 |
| OMTD07 | 29.90 | 30.63 | 42.00 | 34.70 |
| OMTDN07 | 28.93 | 29.72 | 43.20 | 36.93 |

Table 2. Summary of average results for the stemmed three runs

As it can be easily understood from the comparisons of the mean average precision (MAP) scores in Table 2, the title and description run (OMTD07) has achieved the best performance (29.90%), closely followed by the title, description and narrative (OMTDN07) run. Relatively, our title query (OMT07) run has performed slightly worse. When it comes to R-precision scores OMTD07 has again gained the best score followed by OMTDN. It is worth noting that the MAP and R-precision scores we have obtained for these three runs are much better than MAP scores that we had achieved for similar official runs last year. We feel most of these relative improvements in the performances of our CLIR system are resulted from the enhancements and refinements that we have made in the major components of our Oromo-English CLIR system since last year.

As it was indicated earlier, one of the major goals for conducting our experiments was to identify and determine the effects of Afaan Oromo light stemmer on the overall performances of Oromo-English CLIR system. In order to identify the impacts of this light stemmer on our CLIR system, we have conducted and submitted experiments with stemmed and non-stemmed Afaan Oromo queries using all fields from each of the topic sets. Table 3 shows the performance gains caused by Afaan Oromo light stemmer over the baseline run (i.e. NOST-OMTDN07) in terms percentage.

| Run-Id | Stemming | MAP (%) | R-Prec. (%) | P@10 (%) | P@20 (%) |
|---|---|---|---|---|---|
| OMTDN07 | Yes | 28.93 | 29.72 | 43.20 | 33.80 |
| NOST-OMTDN07 | No | 20.38 | 22.17 | 31.40 | 26.00 |
| Change in Percentage | | 41.95 | 34.05 | 37.58 | 30.00 |

Table 3. Comparisons of results of stemmed and non-stemmed runs

The result in Table 3 clearly indicates the significant improvements that have been gained by our light stemmer in relation to the non-stemmed baseline run, i.e. NOST-OMTDN07. Contributing more than 40% MAP score compared to the baseline, the Afaan Oromo light stemmer has much positive effects on performances of our Oromo-English CLIR system. This confirms the findings of earlier similar researches [e.g.10] and suggests that employment of a linguistically motivated rule-based stemmer is very beneficial for development and application of CLIR systems.

Table 4 shows interpolated Recall-Precision scores at the standard eleven recall levels.

| Recall Levels | OMT07 (%) | OMTD07 (%) | OMTDN07 (%) | NOST-OMTDN07 (%) |
|---|---|---|---|---|
| 0.0 | 60.00 | 73.88 | 72.78 | 49.73 |
| 0.1 | 45.15 | 56.61 | 58.24 | 40.92 |
| 0.2 | 39.08 | 50.35 | 49.57 | 34.44 |
| 0.3 | 31.53 | 38.72 | 38.81 | 28.54 |
| 0.4 | 26.93 | 32.94 | 31.40 | 22.59 |
| 0.5 | 23.85 | 28.23 | 27.80 | 18.70 |
| 0.6 | 18.96 | 22.54 | 21.53 | 15.36 |
| 0.7 | 15.83 | 18.56 | 16.95 | 11.54 |
| 0.8 | 11.62 | 13.81 | 12.18 | 8.49 |
| 0.9 | 7.39 | 8.25 | 7.45 | 5.87 |
| 1.0 | 4.60 | 4.33 | 3.70 | 3.34 |

Table 4. Interpolated Recall–Precision scores for the four official runs

As it can be easily observed from the summarized statistics presented in Table 4, our rule-based light stemmer has performed much better than the baseline run almost at all recall levels. This implies the fact that a light stemmer that is designed focusing on major frequent inflectional and attached suffixes of words is very effective and useful in development and application of CLIR for morphologically rich languages like Afaan Oromo.

## 6. Conclusion

The results we have obtained this year in all of our official runs show significant improvement over the last year runs. We feel these relatively good improvements are due to the enhancement of our lexical resources and refinements of the rules of our stemming algorithm. We have also tested and analyzed the major impacts of Afaan Oromo light stemmer on the overall performances of the CLIR system. The application of our light stemmer has significantly outperformed the non-stemmed official run that is used as a baseline in our current experiments. Indeed, it is possible to anticipate such considerable contributions and positive effects of the stemmer since Afaan Oromo is one of the morphologically rich and complex languages.

However, our recent manual analysis and investigation of the individual query results shows the fact that there are gaps in the performances our CLIR system over different individual query topics. Since we are using all translation senses of Afaan Oromo query terms that are found in the bilingual dictionary, certain irrelevant keywords may be included in most of our search queries. Thus, in some instances, the performances of our CLIR system over the individual query topics is much worse than the mean average precision (MAP) score we have achieved in our official runs. We have also observed that some of the query terms are sometimes wrongly over-stemmed or under-stemmed during the suffix removal process which might have led to wrong translation results. We will try to address these and other related important CLIR issues in our future Oromo-English CLIR experiments.

## References


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

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

[3] Kula, K. T., Varma, V. and Pingali, P. Evaluation of Oromo-English Cross-Language Information Retrieval. In *IJCAI 2007 Workshop on CLIA*, Hyderabad (India), 2007.

[4] Gumii Qormaata Afaan Oromoo. Caasluga Afaan Oromoo, Jildii – 1. Komishinii Aadaaf Turizmii Oromiyaa. Finfinnee, Ethiopia, 1995 E.C.

[5] Carpuat, Marin and Fung, Pascale. Simple Dictionary-Based Query Translation. In *CLEF 2001 Working Notes.* 2001.

[6] Kula K. T. and Varma, V. Oromo-English Information Retrieval Experiments at CLEF 2006**.** In *CLEF 2006 Working Notes*, 2006.

[7] Cushitic languages, http://en.wikipedia.org/wiki/Cushitic_languages.

**[**8] Atelach, Alemu, Lars Asker, Rickard Coster, and Jussi Karlgen. Dictionary Based Amharic English Information Retrieval. In *CLEF 2006 Working Notes*, 2006.

[9] Lucene. URL: http://lucene.apache.org.

[10] Larkey, Leah S. Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In *SIGIR'02*, August 11-15, 2002, Tampere, Finland, 2002.