Evaluation of Oromo-English Information Retrieval

Workshop on Cross Lingual Information Access Addressing the Information Need of Multilingual Societies

Kula Kekeba Tune, Vasudeva Verma and Prasad Pingali
(LTRC, IIIT- Hyderabad)

Jan. 6, 2007

Outlines

- Motivations and Contributions
- Overview of CLIR
- Overview of Afaan Oromo
- Related Works
- Experimental Setup
- Evaluation Results
- Conclusions and Future Works

1. Motivations and Contributions

- **Aim**:
  - To design and develop a dictionary based Oromo-English CLIR system with a view to enable Afaan Oromo speakers to search and retrieve relevant documents written in English by using their own native language (Oromo) queries

- **This work is mainly motivated by**:
  - The increasing demand for identifying relevant document across different languages to share and exchange information globally
  - The need for addressing the problem of language barrier by developing and applying CLIR for various languages including indigenous and resource scarce African languages like Afaan Oromo
  - The need for assisting and enabling native speakers of Afaan Oromo to access and retrieve relevant English documents using Oromo queries

1. Motivations and Contributions (contd.)

- **Few of the major contribution of our work include**:
  - Construction and adaptation of basic Afaan Oromo IR tools such as bilingual dictionary, stemmer and stopword list and their applications for Oromo-English CLIR
  - Designing and implementation of the first CLIR system for Afaan Oromo
  - Analysis and review of CLIR research works related to indigenous African languages and sharing of their experiences
  - Testing and assessment of Oromo-English CLIR performance at a standard and internationally recognized evaluation forum like CLEF
  - Demonstrating the feasibility of CLIR application for resource scarce and indigenous African language like Afaan Oromo

CLEF: http://www.clef-campaign.org/

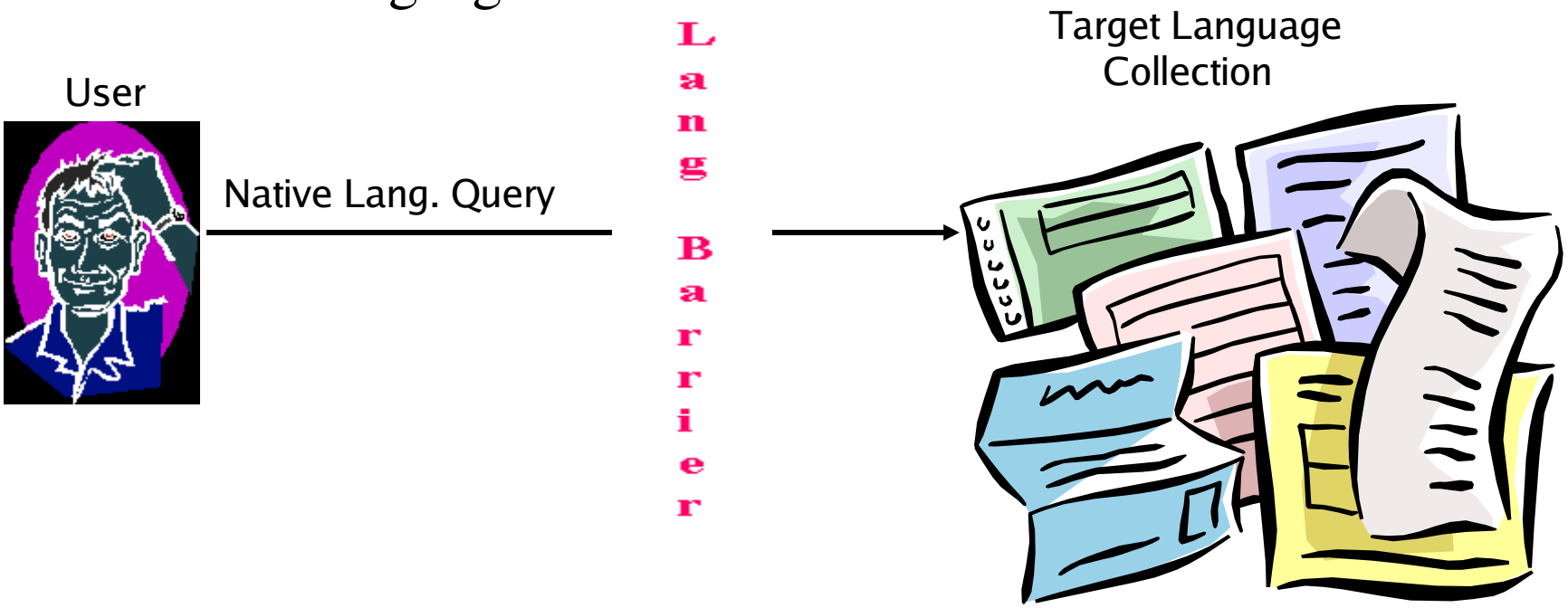
2. Overview of CLIR

2.1 *CLIR* Defined:

- Cross-Language Information Retrieval (CLIR) is a subfield of Information Retrieval (IR) system
- It deals with searching and retrieving information written/recorded in a **language different from the language of the user's query**
- The process is called **bilingual CLIR** when it deals with a language pair, i.e., one source or query language (e.g., Afaan Oromo) and one target or document language (e.g., English)
- And it is called **multilingual CLIR** when it deals with retrieval of documents from multilingual target collections

2.2 Basic Tasks of CLIR

- **Basic Task:** Finding documents of a target language (e.g. English) using queries expressed in user's/source language (e.g. Oromo)
- **Problem:** Language barrier because of documents and queries are in different languages



2.2 Basic Tasks of CLIR (contd.)

- In order to overcome the language barrier **translation** is required:
  - either the query has to be translated into the language of the documents or
  - the documents have to be translated into the language of the query
- Translation of the whole document collection is more demanding than query translation as it requires more scarce resources like full-fledged MT system
- Hence **query translation techniques** has become more feasible and common in developing and applications of CLIR system

2.3 Issues and Approaches in CLIR

- As indicated by Peters and Sheridan (2001), CLIR is a complex multidisciplinary research area in which methodologies and tools developed in the field of information retrieval (IR) and natural language processing converge

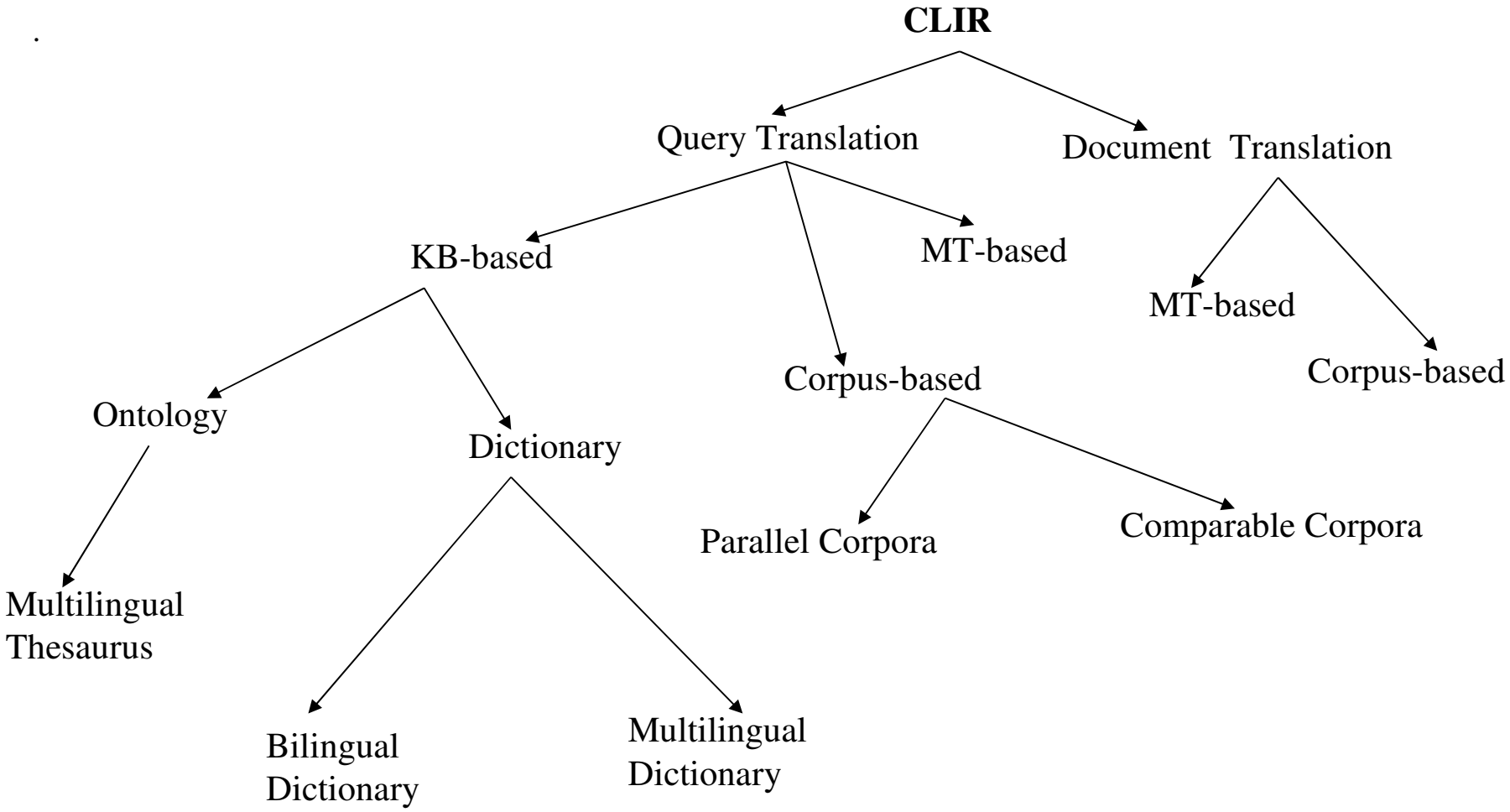
- Some of the major CLIR issues include:

  - What to index?
    - Free text, key words or controlled vocabulary

  - What to translate?
    - Queries or documents

  - How to translate?
    - Using MT, dictionary, ontology or parallel corpua

2.3 CLIR Approaches



3. A Brief Overview of Afaan Oromo

- Afaan Oromo (also known as Oromo) is one of the major Languages that are widely spoken and used in Ethiopia; currently it is an official language of Oromia state
- Unlike Amharic, (an 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*** (Yimam,1986 and Nefa, 1988)
- 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 most of the grammatical features are indicated by affixes
- Both Afaan Oromo nouns and adjectives are highly inflected for number and gender
  - For instance, in comparison to the English regular plural marker ***–s*** (-es), there are more than 12 major and very common plural markers in Afaan Oromo nouns (e.g. ***–oota, oolii, -wwan, -lee, -an, een, eetii, -eeyyi, -ii,*** etc.)

3. A Brief Overview of Afaan Oromo (contd.)

- Surprisingly, a given (single) ***noun or adjective*** can take one or more plural markers in Afaan Oromo
- Oromo noun inflection examples:
  - Plural for markers noun: "**Mana**" (N, House)
    - Mana-wwan
    - Man-oota
    - Man-oolee
    - Mann-een
    - Mann-eetii
    - ***Mann-eetii-wwaan***
  - Gender markers for noun (***-eessa/-eetii, -a/-ttii, -aa/-tuu***, etc.)
  - Examples:
    - Obbol-***eessa*** (M, Brother) vs. Obbol-***eettii*** (F, Sister)
    - Garb-***a*** (M, Servant) vs. Garb-***itti*** (F, Servant)
    - ***Garb + -a/-tti + -oota = Garboota (M, Servants) vs. Garbitoota, (F, Servants)***

3. A Brief Overview of Afaan Oromo (contd.)

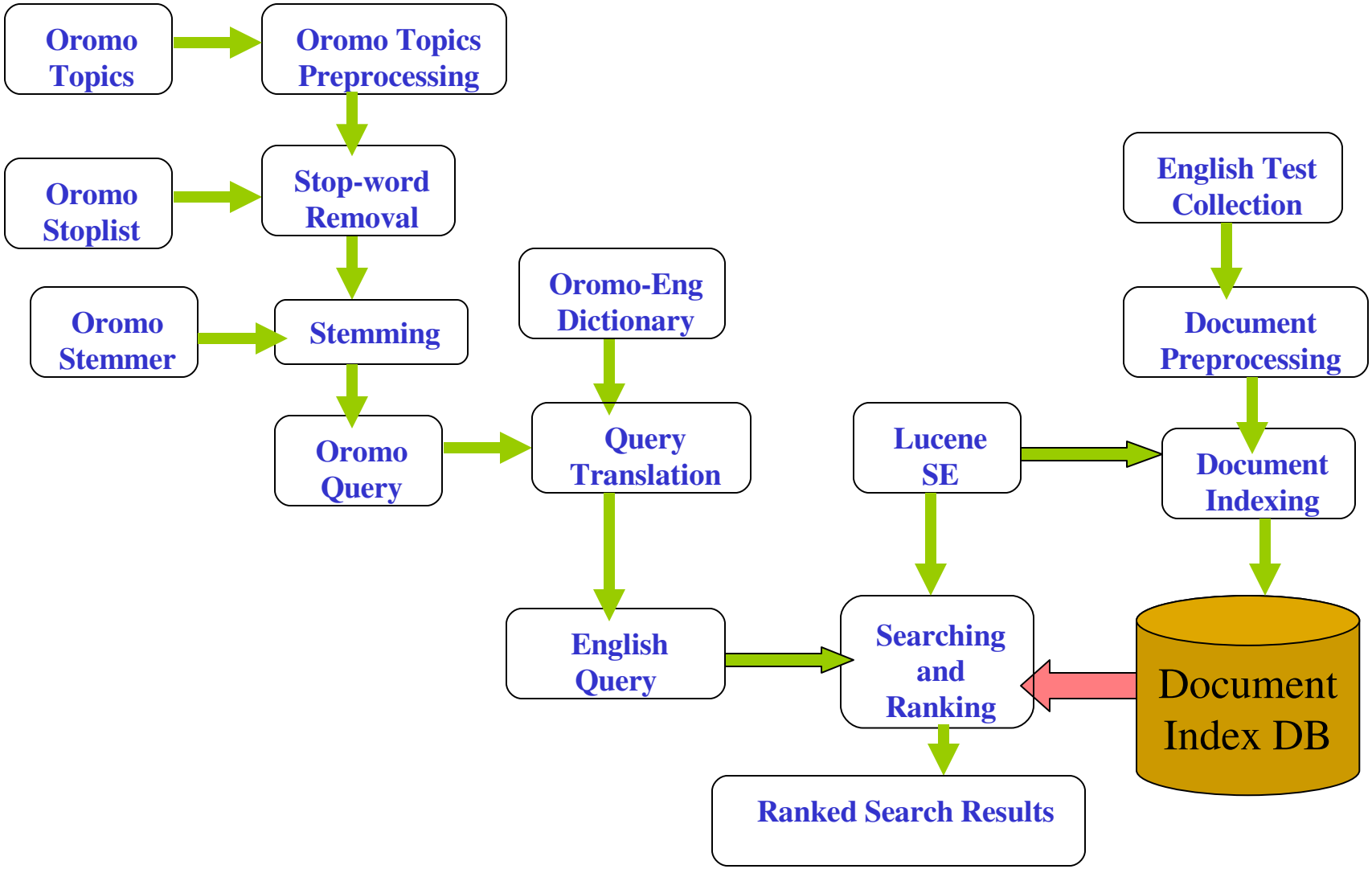
- Afaan Oromo verbs are also highly inflected for gender, person, number and tenses
  - Oromo Verb Inflection Examples:
    - Beek-uu (inf, to know)
    - Beek-a (1st 0r 3rd Singular, M, S.Present)
    - Beek-na (1st Plural, S.Present)
    - Beek-ti (3rd Female, Singular, S.Present)
    - Bebeekan = Be - beek - an (3rd Plural, S.Present, Reduplication)
- Moreover, possessions (***-ko, -ke, -sa***), postpositions (e.g. ***bira, dura, irra, jala***), prepositions (e.g. ***akka, gara, gad***), cases (e.g. ***–n***), auxiliaries (***-dha, -jira***), conjunctions (e.g. ***–fi, -lee, -moo***) and article (e.g.***-icha, -itti***) markers are often indicated through affixes in Afaan Oromo
  - **Examples**: Iskootilaandi***-irra-tti-dha***, Sootroo-***wwan-in-itti-f***
- Morphological derivations and word formations in Afaan Oromo also involve a number of different linguistic features including affixation, reduplication and compounding

4. Related Works

- Very limited works have been done in the past in the areas of IR and CLIR in relation to African indigenous languages including major languages of Ethiopia
- Two CLIR case studies and evaluation experiments were undertaken by (Cosijn et al. 2002 and 2004) for two different major languages in South Africa, i.e. for Zulu-English and Afrikaans-English CLIR
  - A dictionary-based query translation technique was used to translate Zulu and Afrikaans queries into English
  - Afrikaans-English CLIR had achieved a better average precision (MAP) of 19.4%
- Another similar three different dictionary-based CLIR evaluation experiments were conducted on Amharic (another African indigenous and official language of Ethiopia) at a series of CLEF ad hoc tracks (Alemu et al., 2004, 2005, 2006)
- While the two dictionary based Amharic-English CLIR evaluation experiments were conducted at CLEF 2004 and 2006, Amharic-French CLIR experiment was undertaken at CLEF 2005

5. Experimental Setup

5.1. Major Components of Oromo-English CLIR



5.2. Purpose and Contexts of Our CLIR Evaluation

- The purpose of our initial evaluation experiments at CLEF 2006 was to assess the over all performance of the Oromo-English CLIR system by using different fields of Oromo topics
- Thus we submitted to CLEF 2006 three official runs (experiments) that differed in terms of utilized fields in the topic set, i.e. title run (OMT), title and description run (OMTD), and title, description and narration run (OMTDN)

Collection	Number of Docs	Size of Docs (MB)
LA Times-94	113,005	425
GH Herald-95	56,472	154
Total	169,477	579

Table 1. Summary of English Test Collection

- All of these three experiments have been carried out by using standard CLIR evaluation resources provided by CLEF for ad hoc track bilingual tasks
- ***Lucene***, which is an open source search engine that is mainly based on vector space model was adopted and used for indexing and retrieval of the English documents

5.3 Afaan Oromo Stopword Lists and Stemmer

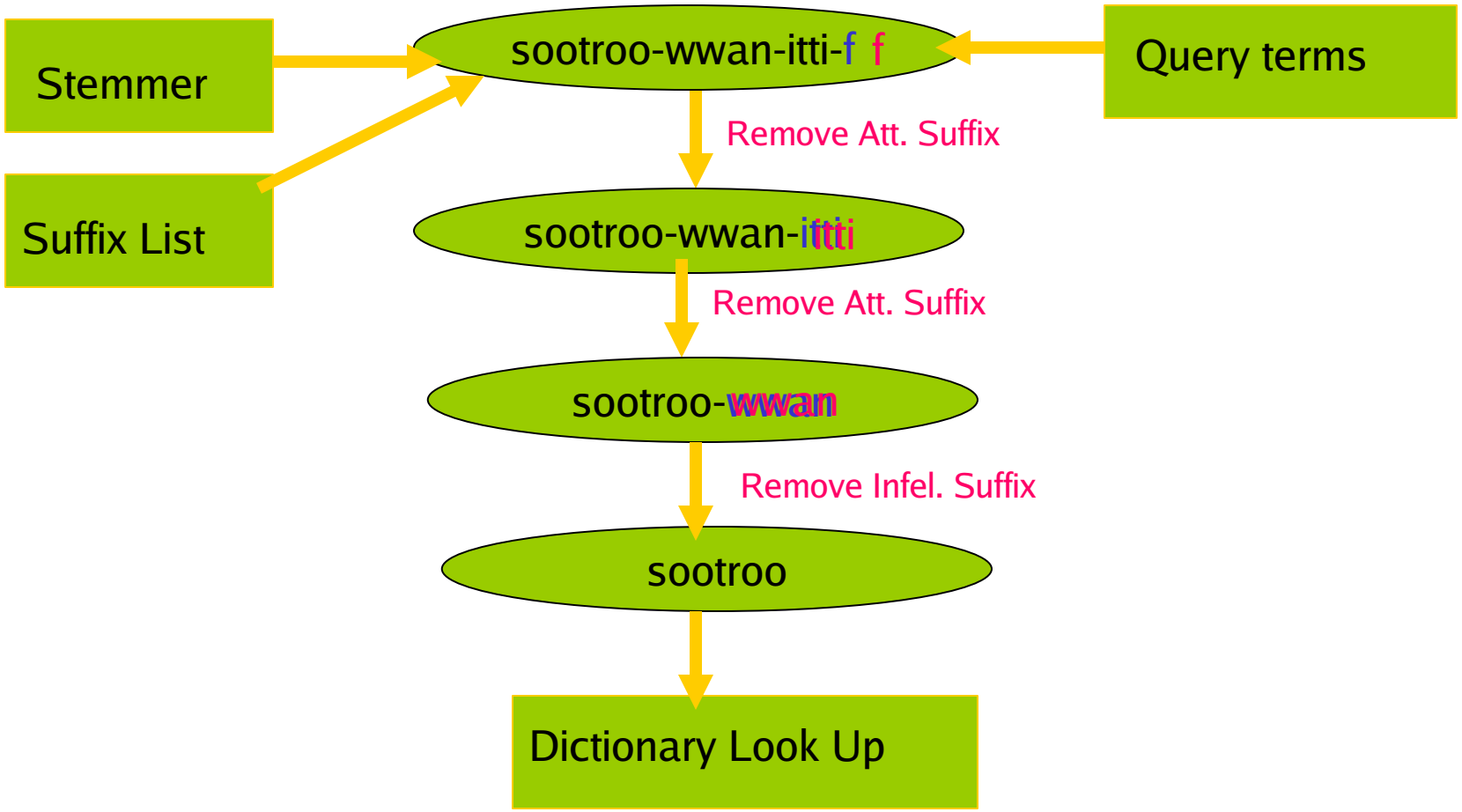
- In order to define Oromo stopwords, we first generated and created a list of the top 350 most frequent words found in 1.2 million words of Afaan Oromo text corpus by using TF/IDF measures
- Then we incorporated additional pronouns, conjunctions, prepositions and other similar functional words in Afaan Oromo and used about 580 stopwords in conducting our experiments
- Once these stopwords were removed from the Afaan Oromo topics, we applied a light stemming algorithm in order to conflate word variants into the same stem or root
- As a number of previous research works on CLIR (including Carpuat and Fung, 2001) have indicated languages that are morphologically rich can benefit a lot from stemming
- Since Afaan Oromo is morphologically very rich and stemming is often language dependent, we have developed a rule based suffix-stripping algorithms focusing on very common inflectional suffixes of Oromo words

5.3 Afaan Oromo Stemmer (contd.)

- Some of the common suffixes that have considered in our current light stemmer include gender (masculine, feminine), number (singular or plural), cases (nominative, dative), post-positions, prepositions and possession in Afaan Oromo
- Broadly speaking, it is possible to categorize suffixes in Afaan Oromo into three basic groups:
  - Derivational suffixes for noun, verbs and adjectives (***e.g. –eenya, -ummaa***)
  - Inflectional suffixes for number and gender (e.g. ***ii, –lee, –oota, -te, -wwan***)
  - Attached suffixes such as postpositions (e.g. ***-arra, -bira, -irra, -itti, -dha***)
- Based on our current observations, the most common order/sequence of Afaan Oromo suffixes in a given word (right to left) is ***derivational, inflectional and attached*** suffixes
- Thus, the stemmer is expected to remove from the right end first all possible attached suffixes, then inflectional suffixes and finally derivational suffixes (if necessary)

5.3 Major Steps of Stemming (contd.)

- The following simple query term stemming example illustrate some of the major stemming procedures



5.4. Query Translation

- Topics (queries) in CLEF ad hoc track are structured statements representing user's information needs
- Each topic consists of **three parts**: a brief "***title***" statement; a one-sentence "***description***"; a more complex "***narration***" often specifying the relevance/irrelevance of a document
- Sets of 50 topics were prepared for the ad hoc of CLEF 2006 bilingual tasks for which a participant is expected to retrieve top 1000 documents for each and every ***query*** submitted for the official run
- Example of Afaan Oromo topic from CLEF 2006:

<top>
<num> C308 </num>
<OM-title> Gaaddiddeeffamuu Aduu </OM-title>
<OM-desc> Dokumantoota guutumaan gaaddiddeeffamuu ykn cinaan gaaddiddeeffamuu aduu gabaasan barbaadi. </OM-desc>
<OM-narr> Dokumantootni gaaddiddeeffamuu aduu irratti odeeffannoo kamiyyuu kennan fudhatama ni qabu. Dokumantootni waayee gaaddiddeeffamuu baatii ykn sosochiiwwan pilaaneetootaa ibsan as keessa hin galan. </OM-narr>
</top>

5.4. Query Translation (contd.)

- In order to translate Afaan Oromo topics into bags-of-words of English queries, we have used Oromo-English dictionary which was adopted and developed from hard copies of human readable bilingual dictionaries by using OCR technology
- After stemming, the query terms of Oromo topics were automatically looked up for all possible translations in this bilingual dictionary
- Therefore, the resulting English queries were simple bags-of-words, taking into account all possible translation of Oromo query keywords found in the bilingual dictionary
- One of the major problems in this translation process was related to handling out of dictionary or unmatched query words most of which are **proper names** like:
  - *Xaaliyaanii, Kurdii and Buush*, and
  - **foreign or borrowed words** like:
    - *fiilmi* and *oopiyeemii*

6. Experimental Results

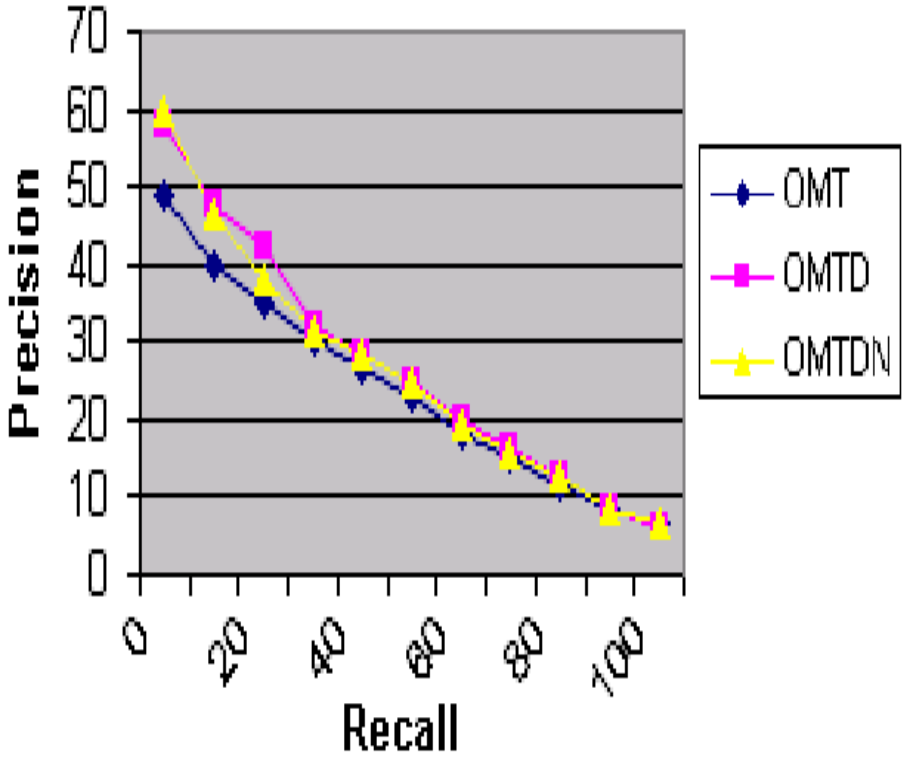
- The Mean Average Precision (MAP) scores, number of Relevant-total, Relevant Retrieved and R-Precision of our three runs (OMT, OMTD, OMTDN) are summarized and presented in the frist Table
- The second Table shows summary of Recall-Precision results for the three runs

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

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

6. Experimental Results (contd.)

- As it can be observed from the first Table, there is no significant difference between the MAP of the three runs though the title run has slightly lower performance with MAP of 22% which might be due to the fact that most of the title fields are very short
- The OMTD (title and description) run has achieved the best performance (with MAP of 25.04%) in our current experiments



- The Precision-Recall curve depicted above further illustrates the performance of our CLIR system at different recall levels

7. Conclusions

- In this paper we have tried to describe the basic components and features of our experimental Oromo-English CLIR system together with its official evaluation results at CLEF 2006
- Based on this dictionary-based CLIR experiments we have attempted to show how very limited language resources such as bilingual dictionaries and light stemmers can be used in a standard information retrieval evaluation setting
- Since this is the first time we participated in CLEF campaign, we feel we have obtained reasonable average results for all of our official runs, given the limited resources and simple approaches that we have used in our CLIR experiments
- There is a growing demand for development and application of CLIR in a number of indigenous and resource scarce African languages
- Thus, we feel our results will encourage other researches to design and develop similar CLIR system for these major indigenous languages despite they have very limited linguistic resources and IR facilities

8. Future Works

- There are lots of rooms for improvement of the performance of our Oromo-English CLIR systems. Some the remaining important tasks include:
  - Evaluation of the impacts of different resources and components of our CLIR system
  - Query expansion by using relevance feedback and related techniques
  - Handling of out of dictionary words (like proper names and foreign terms)
  - Application of some crude disambiguation mechanisms
- The task of Oromo phrasal terms identification and compound words handlings are also the other important research issues in order to improve the performance of the Oromo-English CLIR system

THANK YOU!