

CLEF 2005: Multilingual Retrieval and Merging (MuRaM) Track

Call for Participation

Gareth J. F. Jones
Dublin City University

1 Why Multilingual Retrieval and Merging?

A longstanding, and as yet unsolved, research topic in multilingual information retrieval is the effective formation of a single merged output list combining documents from the available collections in different languages. The merging strategies explored previously for multilingual retrieval tasks at CLEF and elsewhere have generally produced disappointing results. These multilingual tasks have required the participants to carry out retrieval from document collections in various different languages and perform the merging operation. This means that participants have often spent much of their effort focussing on developing retrieval systems for each document language, often leaving little opportunity to consider the issues of effective merging in detail. Specifying the multilingual retrieval task in this way has also meant that it is only open to those researchers able or willing to build retrieval systems for each of the document languages.

The CLEF 2005 Multilingual Retrieval and Merging Track (MuRaM) comprises two subtasks: a traditional multilingual retrieval task requiring participants to carry out retrieval and merging, and a new task focussing on multilingual merging from provided standard sets of ranked retrieval output. In running this new multilingual task, it is our hope that participants will be able to explore the issues involved in multilingual merging in detail and propose novel solutions to this problem, and that the merging only task will encourage participation by researchers interested in exploring the merging problem without the need to build retrieval systems for the document languages. The use of common retrieval data sets will also enable direct comparison of the behaviour and performance of the proposed merging techniques independent of trying to take into account the differing underlying ranked lists generated, as has been the case in previous evaluations.

2 Task Details

The CLEF 2005 Multilingual Retrieval and Merging task will be based on the CLEF 2003 Multilingual-8 (Multi-8) task. The target document collections for this task are in the following 8 languages: Dutch, English, Finnish, French, German, Italian, Spanish, and Swedish. Thus, unlike previous multilingual retrieval tasks which have generally required participants to develop retrieval systems for new document languages, focusing on languages for which retrieval systems have already been developed will hopefully enable participants to focus on merging issues and/or improving retrieval for existing document languages. The original CLEF 2003 Multi-8 task includes 60 search topics and relevance assessments. For the CLEF 2005 multilingual task will use these topics with the original relevance assessments. However, for CLEF 2005 the topic set will be divided with 20 topics assigned as a development set and the remaining 40 topics used for the evaluation runs.

Two levels of participation are available:

- **Multi-8 Two-Years-On:** This task offers participants the opportunity to carry out the original 2003 Multi-8 task by performing their own retrieval runs and submitting their merged multilingual results. Participants are required to return at least one run for English topics using the Title and Description topic fields. Additional runs may be submitted using alternative topic fields. Topics are available in a large number of languages, and participants are also invited to return additional runs using topics in one or more of these other languages. A standard set of “English - target language” topics translations will be available to participants wishing to make use of these. We would be interested to hear from anyone intending to participate in this subtask who would be willing to contribute the outputs of their individual monolingual runs for use in the Merging Only task.
- **Multi-8 Merging Only:** Participants will investigate merging algorithms using provided ranked lists. These will be generated using English search topics using the Title and Description fields of the topics. The minimum requirement for participants is to submit one run for each set of lists using only the information in the standard TREC ranked lists. This submission is intended as a baseline representing what can be achieved using information available to existing merging strategies. Beyond this participants may use any combination of the information in the extended forms of the ranked lists provided described below. Ranked lists will be generated using various retrieval approaches. Further details of the list sources will be provided with the lists. Original documents will also be available to participants wishing to make use of these within their merging strategy.

Further details of the task and the results of original runs submitted for CLEF 2003 can be found from the CLEF workshop website at: www.clef-campaign.org.

Important dates:

Registration Opens	31 January 2005
Data and Topic Release (Multi-8 Two-Years-On)	15 February 2005
Development Ranked Lists Release	31st March 2005
Test Ranked Lists Release	30th April 2005
Submission of Runs	30th June 2005
Results Released	15 July 2005

2.1 Rationale

List merging schemes vary in the information from the individual lists used in the merging process to decide the overall merged list rank. Most existing schemes used for multilingual fusion are fairly simple. Some examples of typical merging schemes are: “round robin” which takes one item from the top of each list in a cycle, round robin extended to take documents from the top of the list in ratio to the collection size, raw score merging which uses the matching scores from each ranked list to interleave the documents in the merged list, and various normalized forms of raw score merging.

Generating merged lists in this way requires very little information in the ranked list. In the simplest case of round robin only the identity of each document and its relative position in the ranked list is required. The more complicated schemes require the matching score for each document and the collection size. However, existing merging schemes have not been found to produce very good merged multilingual results.

In order to enable participants in CLEF 2005 multilingual merging task to approach multilingual merging more creatively the ranked retrieval lists will include a number of features generated in the matching process.

The utility for effectively using these factors in list merging is unknown. They are provided to enable participants to develop and explore new merging strategies. Participants are free to use as few or as many of the available pieces of information as they like.

2.2 Format

The format for ranked lists is based on an extended version of the TREC standard submission format.

The standard TREC submission format adopts the following standardized format:

```
030  00  ZF08-175-870  0  4238  prise1
qid  iter      docno      rank  sim  run_id
```

where

```
qid      is the query id
00       is ignored
docno    is the document id
rank     is the rank of the document
sim      is the matching score for this document
run_id   is a unique identifier for this run.
```

While the data available from ranked lists in this format is sufficient for simple merging schemes, it restricts possibilities for exploration of new schemes.

The following is the proposed extended format:

```
550765  35.5  9  3.5
N      adl  R  fbw

030      00      ZF08-175-870  0  4238  prise1  27  3
qid      iter      docno      rank  sim  run_id  dl  noterms

documen  565      2  8  retriev  760  2  5
term1    n1      tf1      r1  term2    n2    tf2  r2

...      10      system  350  2  1  inform  104
...      noeterms  eterm1  en1  etf1    er1  eterm2  en2

2  4  ...
etf2  er2  ...
```

The new features of this list are defined as follows.

N	is the number of documents in the collection
adl	is the average document length in terms
R	is the number of relevant documents in any feedback process
fbw	is the relative weighting of original and feedback topic terms
dl	is the document length of document docno
noterms	is the number of terms in the original topic statement matching with this document
termi	is the <i>i</i> th matching term in the topic statement
ni	is the number of postings of the <i>i</i> th term in the document collection
tfi	is the term frequency of the <i>i</i> th term in document docno
ri	no of relevant documents containing the <i>i</i> th term
noeterms	is the number of feedback expansion added in any feedback process that match in this document
etermi	is the <i>i</i> th matching expansion term in the topic statement
eni	is the number of postings of the <i>i</i> th expansion term in the document collection
etfi	is the term frequency of the <i>i</i> th expansion term for document docno
eri	no of relevant documents containing the <i>i</i> th expansion term

This additional information should make it possible to compute a new matching score for each document based on the collection parameters. If the details of the original retrieval system are available, the original score can probably be reconstructed to a reasonable approximation.

However, the real intention of making all this information available to the merging process is to enable use of cross-language linguistic resources or the exploration of the application of more sophisticated techniques from monolingual distributed information retrieval.

Provided lists will include such extended fields as the donators of the standard lists can easily provide.

3 Registration

You can register for the CLEF 2005 MuRaM track by contacting Carol Peters (carol.peters@isti.cnr.it), the main co-ordinator of CLEF. For further information about the MuRaM track, please contact Gareth Jones (Gareth.Jones@computing.dcu.ie).