The Impact of Evaluation on Multilingual Text Retrieval

Julio Gonzalo
UNED
c/Juan del Rosal, 16, 28040 Madrid, Spain
+34 91 398 7922
julio@lsi.uned.es

Carol Peters
ISTI-CNR
Via Moruzzi, 1, 56124 Pisa, Italy
+39 050 3152897
carol.peters@isti.cnr.it

ABSTRACT
We summarize the impact of the first five years of activity of the Cross-Language Evaluation Forum (CLEF) on multilingual text retrieval system performance and show how the CLEF evaluation campaigns have contributed to advances in the state-of-the-art.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Search process;
H.3.4 [Systems and Software]: Performance evaluation

General Terms
Measurement, Performance, Experimentation, Languages

Keywords
Multilingual Information Access, System Evaluation.

1. INTRODUCTION
The objective of the Cross-Language Evaluation Forum (CLEF) is to promote research into systems for all kinds of multilingual information access through the creation of reusable test-collections and the organization of annual evaluation campaigns. The aim is to encourage the development of user-friendly multilingual, multimedia systems. In the first years, the core track - also known as the ad-hoc track – has consisted of a number of monolingual, bilingual and multilingual document retrieval tasks and has produced the largest available testbed for the study of Cross-Language Text Retrieval. We summarize the features of this testbed and the scientific results of five years of comparative evaluation campaigns which have created this data set.

The CLEF strategy has been to help groups to work their way up gradually from mono- to true multilingual text retrieval by providing them with facilities to test and compare search and access techniques over many languages, pushing them to investigate the issues involved in processing a growing number of languages with different characteristics. Over the years the language combinations have increased and the tasks offered have grown in complexity until, in CLEF 2003, the multilingual track included a task which entailed searching a collection in 8 languages, selected to cover a range of language typologies and linguistic features (see Table 1).

3. CLEF Test Collection
The criteria normally adopted to create a test collection, consisting of suitable documents, sample queries and relevance assessments, have been adapted in CLEF to satisfy the particular requirements of the multilingual context. All language dependent tasks such as topic creation and relevance judgments are performed in a distributed setting by native speakers.

Table 1. CLEF 2000 – 2004 Ad-Hoc Track

<table>
<thead>
<tr>
<th>Collection</th>
<th>Monolingual</th>
<th>Bilingual</th>
<th>Multilingual</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE;FR;IT</td>
<td>X→EN</td>
<td>X→DE,EN,FR,IT</td>
<td></td>
</tr>
<tr>
<td>DE;ES;FR;IT NL</td>
<td>X→EN, X→NL</td>
<td>X→DE,EN,ES,FR,IT</td>
<td></td>
</tr>
<tr>
<td>DE;ES;FI;FR IT;NL;SV</td>
<td>X→DE;ES;FI;FR</td>
<td>X→DE,EN,ES,FR,IT</td>
<td></td>
</tr>
<tr>
<td>DE;ES;FI;FR IT;NL;RU;SV</td>
<td>IT→ES;DE→IT FR→NL;FI→DE; X→RU; X→EN</td>
<td>X→DE,EN,ES,FR,IT</td>
<td></td>
</tr>
<tr>
<td>FI;FR;RU;PT</td>
<td>ES/FR/IT/RU→FI DE/FI/NL/SV→FR X→RU; X→PT</td>
<td>X→FI,FR,RU,PT</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 gives some idea of the dimensions of the CLEF adhoc track testbed. Fuller and more precise details can be found in [1]. Great care has been taken to ensure that the test collection is reusable. Each year the reliability of the assessments pool is tested [2]. The statistical significance of differences has been found to be very similar to those of TREC and usually not influential.
4. RESULTS

The results in terms of different approaches and techniques tested have been impressive. Groups submitting results over several years have shown improvements in performance and flexibility in advancing to more complex tasks. Much work has been done on fine-tuning for individual languages, while other efforts have concentrated on developing language-independent strategies.

4.1 Performance Improvement

As test collections and tasks vary over years it is not easy to document improvements in system performance. For bilingual retrieval evaluation, a common method is to compare results against monolingual baselines. We give the following indications:

In 1997, at TREC-6, the best CLIR systems had the following results:
- EN → FR: 49% of best monolingual French IR system
- EN → DE: 64% of best monolingual German IR system

In 2002, at CLEF, where there was no restriction on topic and target language the best systems gave:
- EN → FR: 83.4% of best monolingual French IR system
- EN → DE: 85.6% of best monolingual German IR system

However, CLEF 2003 enforced the use of “unusual” language pairs, with the following impressive results:
- IT → ES: 83% of best monolingual Spanish IR system
- DE → IT: 87% of best monolingual Italian IR system
- FR → NL: 82% of best monolingual Dutch IR system

With respect to multilingual retrieval (L1 → Ln), the number of systems participating in this task has grown steadily, peaking to a high of 14 groups in CLEF 2003. Similarly to TREC, over the years, we have seen a convergence of techniques and results with very little statistical difference between the best systems. The know-how collectively developed by CLEF participants include language-specific resources (stemmers, stop-word lists, software to handle compounds in agglutinative languages, web-mined parallel corpora), algorithms to merge results from different languages (e.g. the 2-step RSV originally introduced at CLEF [3]), or strategies to index agglutinative languages without linguistic knowledge (such as n-gram character indexing).

4.2 Lessons Learned

The first lesson is that there is no magic recipe or single best algorithm to build an optimal MLIR system. The best system, as proved by CLEF evaluations, is always the result of careful tuning of every system component, and of combining different algorithms and information sources for every MLIR subtask [5]. These are a few issues for which CLEF evaluations have provided qualitative and quantitative evidence along the years – details can be found in the CLEF publications at [4].

Tuning of document and query indexing to individual languages is essential to achieve optimal performance. This includes language-specific stemmers, tf.idf weights and – most notably - strategies to handle decompounding for agglutinative languages. Linguistic tools and resources help to reach top performance, as evidenced by the fact that the best monolingual systems tend to be developed by research groups with native knowledge of the language and specific linguistic tools. In the absence of linguistic knowledge, however, n-gram character indexing has proved an efficient and robust way of handling agglutinative languages [6].

The best strategy for query translation is to combine evidence from multiple sources, including bilingual dictionaries, MT systems and translation probabilities derived from parallel corpora (also automatically mined Web corpora). Document translation can be successfully exploited together with query translation, but the performance gain does not seem to compensate for the extra computational effort. Interestingly, in absence of any translation resource (highly probable for certain language pairs), there are strategies to obtain reasonable baseline CLIR (such as unified character n-gram indexing for close language pairs).

Remarkably, explicit resolution of translation ambiguity (which seemed the most crucial CLIR problem) is positive but not indispensable to obtain competitive results. Black-box MT systems, raw (unigram) translation probabilities, or Pirkola’s structured query approach (where alternative translations are treated as synonyms) might suffice, provided that all other system components are well tuned. Preserving ambiguity is, in any case, far better than filtering translation alternatives poorly.

Besides indexing and translation, the multilingual retrieval problem requires merging results from different languages. Trivial merging strategies, such as raw-score merging or round-robin techniques, provide reasonable results on the CLEF testbed, due to the balanced amount of relevant information across languages. Strategies specifically tailored to the multilingual problem – such as 2-step RSV - provide, however, substantial benefits for the multilingual retrieval problem.

The CLEF evaluation campaigns are, probably, the largest and most comprehensive research initiative for CLIR, but they are far from being complete. One reason is that ad-hoc CLEF tasks evaluate CLIR systems as a whole; there is no direct comparison of alternative solutions for specific system components, such as translation strategies (given a fixed set of translation resources), or resource acquisition techniques (given a fixed translation strategy). This problem has started to be addressed in CLEF 2005, with a task devoted specifically to evaluating merging strategies. Furthermore, in order to meet the needs of applications, over the years, the focus of CLEF has shifted from ad-hoc document retrieval to address other CLIR problems such as interactive issues, multimedia retrieval, cross-language Q&A and now, in 2005, cross-language Web and Geographic Information retrieval.

5. REFERENCES