Adding Multilingual Information Access to the European Library

Martin Braschler[1] and Nicola Ferro[2]

[1] Zurich University of Applied Sciences - Switzerland
martin.braschler@zhaw.ch
[2] University of Padova - Italy
nicola.ferro@unipd.it

Abstract. A feasibility study was conducted within the confines of the DELOS Network of Excellence with the aim of investigating possible approaches to extend The European Library (TEL) with multilingual information access, i.e. the ability to use queries in one language to retrieve items in different languages. TEL uses a loose coupling of different search systems, and deals with very short information items. We address these two characteristics with two different approaches: the "isolated query translation" approach, and the "pseudo-translated expanded records" approach. The former approach has been studied together with its implications on the user interface, while the latter approach has been evaluated using a test collection of over 150,000 records from the TEL central index. We find that both approaches address the specific characteristics of TEL well, and that there is considerable potential for a combination of the two alternatives.

1 Introduction

This paper reports on a feasibility study ([1], [6], [12]) conducted in collaboration between DELOS, the European Network of Excellence on Digital Libraries, and The European Library (TEL). TEL is a service fully funded by the participant members (national libraries) of the Conference of European National Librarians (CENL). It aims at providing a co-operative framework for integrated access to the major collections of the European national libraries. The study intends to provide a solid basis for the integration of multilingual information access into TEL.

By multilingual information access (MLIA) we denote search on collections of information items (in the context of this paper, bibliographic records) that are potentially stored in multiple languages. The aim is to allow the user to query the collection across languages, i.e. retrieving information items not formulated in the query language. The term "cross-language information retrieval" (CLIR) is often used to describe this definition of MLIA, distinguishing it from monolingual access to information in multiple languages (which is already implemented in TEL).

Today, mainstream research on CLIR in Europe is carried out within the confines of the *Cross-Language Evaluation Forum* (CLEF) campaign [14]. Most of the experiments in CLEF concentrate on retrieval of lengthy, unstructured full-text

C. Thanos, F. Borri, and L. Candela (Eds.): Digital Libraries: R&D, LNCS 4877, pp. 218–227, 2007.

documents using a general vocabulary. An overview of the recent achievements in CLIR can be found in [5], [10], and [11]. Generally, there is a growing sense among the academic community that the CLIR problem as applied to such lengthy, unstructured full-text documents from a general domain is fairly well understood from an academic standpoint [3], [4]. Unfortunately, the situation in the TEL system is substantially different from the ideal "mainstream setting" for CLIR. TEL employs only a loose coupling of systems, i.e. each query is forwarded to the individual libraries. In such cases, translation and retrieval cannot be tightly integrated. We address this problem with the "isolated query translation approach" (Section 3). Furthermore, the large majority of information items are very short. Similarly, the expressions of information needs by the users, i.e. the queries, tend to be very short as well (average length is 2.2 words). These contradictions to the general CLIR setting are addressed by our "pseudo-translation on expanded records" approach (Section 4).

2 TEL Architecture and Functioning

Figure 1 shows the architecture of the TEL system. The TEL system allows easy integration of national libraries [15] by extensively using the *Search/Retrieve via URL* (SRU)[1] protocol. In this way, the user client can be a simple web browser, which exploits SRU as a means for uniformly accessing national libraries.

With this objective in mind, TEL is constituted by three components: (1) a Web server which provides users with the TEL portal and provides integrated access to the national libaries via SRU; (2) a "central index" which harvests catalogue records from national libraries which support the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH)[2] ; (3) a gateway between SRU and Z39.50[3] which allows national libraries that support only Z39.50 to be accessible via SRU.

This setup directly influences how MLIA/CLIR can be integrated into TEL. Indeed, the TEL system has no control on queries sent to the national libraries, as interaction with national library systems is via SRU. Consequently, introducing MLIA functionalities into the TEL system would have no effect on unmodified national library systems. Modification of these systems, however, is an unviable option due to the effort required and the "low barrier of entry" criteria adopted when designing the TEL system.

Therefore, while still offering some MLIA functionalities, we have investigated the possibility of adding an "isolated query translation" step. Additionally, the TEL "central index" harvests catalogue records from national libraries, containing catalogue metadata and other information useful for applying MLIA techniques, such as an abstract. We show how to extend the functionality of this central index to MLIA by adding a component that pseudo-translates the catalogue records ("pseudo-translation of expanded records"). This addresses the brevity of the information items involved, and is substantially different from approaches on the ideal "mainstream setting" for CLIR.

[1] http://www.loc.gov/standards/sru/

[2] http://www.openarchives.org/OAI/openarchivesprotocol.html

[3] http://www.loc.gov/z3950/agency/

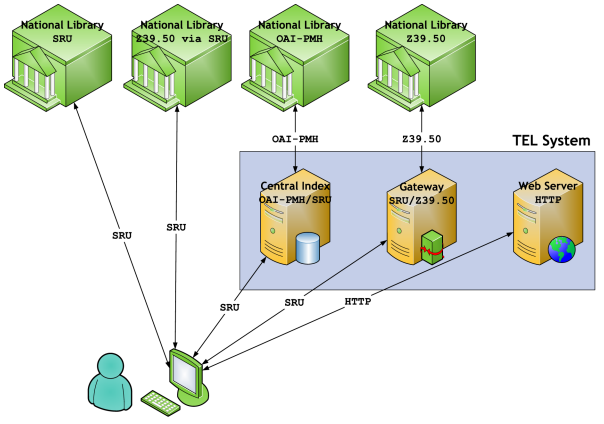

Fig. 1. Present architecture of the TEL system

3 Isolated Query Translation

"Isolated Query Translation" addresses the problems for MLIA generated by the loose coupling of the systems of the individual national libraries. The new component can be directly accessed by the client browser using the SRU protocol. It can be considered as a sort of pre-processing step where the translation problem is treated as completely separate from retrieval.

The approach works as follows: (1) before actually submitting the query, the user asks the browser to translate it; (2) the browser sends the query via SRU to the "isolated query translation" component, which takes care of translating it and, if necessary, applies query expansion techniques to reduce the problem of missing translations; (3) at this point, the user can interactively select the translation which best matches his needs or can change some query term to refine the translation. In this latter case, the translation process may be iterated. (4) Once the desired translation of the query has been obtained, the retrieval process is initiated, using both the translated query and the original one.

This solution is easy to implement and complies with the "low barrier of entry" approach. The national library systems do not require any modification and this new functionality can be transparently applied when querying them. Some user interaction is required, because multiple translations of the same term may need to be disambiguated or the original query may need to be modified.

The main drawback of this approach the separation of the translation from the retrieval process. Relevant documents may be missing in the result set and thus the performance can be low. Moreover, huge linguistic resources, such as dictionaries, are needed since the vocabulary used in queries is expected to be very large; this has to be repeated for each pair of source/target language the system is going to support. Finally, the query expansion mechanism has to be generic and cannot be tailored on the collections queried, since the "isolated query translation" component does not interact with the national library systems.

3.1 Modifications to the TEL System User Interface

In our discussion on how to modify the current user interface of the TEL system for the "isolated query translation" feature we focus our attention on the simple search functionality. First, the interface is extended with an additional link to the "Isolated Query Translation" feature. When the user clicks on the "suggest query in other languages" link (Figure 2), a box with the supported target languages for the translation appears below the search input box. The user can now check the languages for which he wants a translation of the query.

Fig. 2. Selection of the source and target languages in the simple search

Moreover, on the left of the search input box, a list with the possible source languages of the query is now shown, so that the user can specify the language of his original query. Note that the set of languages for the user interace may differ from the languages available for translation.

As shown in Figure 3, for each target language selected by the user, a new text input box appears below the search input box containing the translation of the query in that language. There are different possibilities for managing the user interaction when the translation of the query is shown. A first possibility would be to add a button "Suggest" so that the user presses it and the input boxes with the translation of the query appear (explicit request by user). Another possibility would be a more *Asynchronous JavaScript Technology and XML* (AJAX) style of interaction where the input box with the translation appears as soon as the user selects the target language. In any case, both ways of interaction comply with the current approach of the TEL system in developing the user interface, which already exploits AJAX.

Once the translations have been obtained, we want to allow the user to modify or refine the translations. Since the users of the simple search probably prefer an easy and intuitive way of interacting with the system, the translation refinement step should also be as simple as possible, even if some precision or some expressive power is lost. For this purpose, we can assist the user as follows:

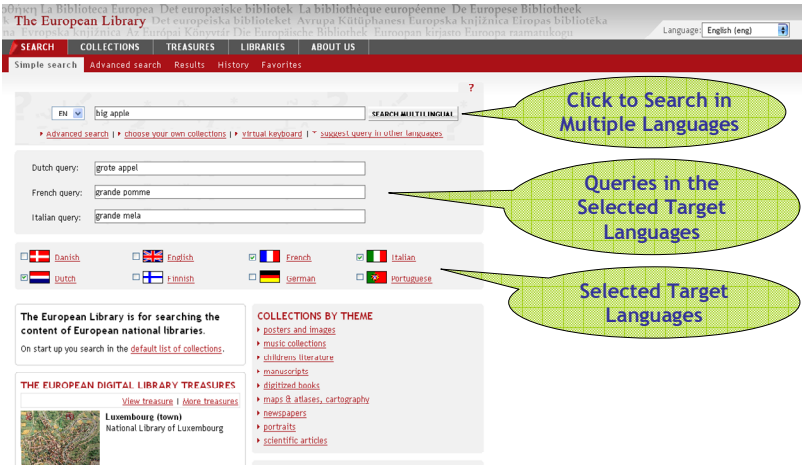

Fig. 3. Query suggestions in other languages in the simple search

1. The user could directly edit each text input box. This means that the user can delete or add words to the translation. If the user has no knowledge of a target language, he will have to use the query suggested by the system without modifications.
2. Some attention has to be paid when multiple translations are possible because they all have to be listed in the input box and thus some visual clue should be provided to help the user in distinguishing between multiple alternatives.
3. If the translation greatly differs from users expectations, there is the possibility of modifying the source query by adding or deleting words to it, thus obtaining a new translation in the target languages.

Once the various translations of the query have been approved, the user can click the "Search Multilingual" button to perform a search in both the original and the selected target languages.

4 Pseudo-Translation of Expanded Records

Today, the large majority of all records available through the search facility in TEL contain bibliographical metadata only (no abstracts or full text). Only short segments of text are thus available for free-text search by the user (such as the "title" field).

While potentially a problem in monolingual search as well, the brevity of the available text exacerbates the problems usually encountered in multilingual information access situations.

4.1 Expansion Techniques and Pseudo-Translation

The solution that was chosen for overcoming the lack of textual content in the information items is automatic expansion of the content fields. The approach used is derived from techniques used in classical information retrieval for query expansion,

such as relevance feedback [13]. These techniques extract new statistically related search terms from items that are ranked highly in initial searches. While often used interactively, involving a user picking relevant items from a result set, these techniques can also be applied in an automated fashion. In such cases, the system assumes that the top items returned in response to user requests are relevant. Such a technique is called "blind feedback", and has proven to be beneficial in many CLIR settings [2].

It is possible to use the same techniques independently of specific information needs, by expanding the information item itself instead of the query. While usually not applicable to retrieval on lengthy documents, we expected potential for such an approach in the case of the very short records present in the TEL collection. By using expansion techniques, we intended to address both problems of vocabulary coverage and word sense ambiguity, as usually experienced during translation. Additional terms added during expansion tend to be from a more general vocabulary, as term frequency influences term selection. The new, longer representation of the record also makes it less likely that none of the terms can be translated.

In this proposed solution, we cross the language boundary by translating the "document", i.e. the complete record. Document translation has been found to be very competitive [7] in some general cross-language retrieval settings, although query translation is more prevalent. The main reason for the scarce adoption of the document translation techniques can be attributed to problems of scalability. This problem is much less pronounced in the case of TEL, where the brevity of the records should make document translation applicable even to large numbers of records, e.g. in the order of multiple millions of records. The approach is mainly suitable for integration with the TEL central index. However, the same approach could also be deployed in additional search systems of the national libraries that are accessed remotely via TEL.

Using the translated records for matching with queries only, and not for presentation, means that we can use "pseudo-translations", i.e. to potentially leave terms untranslated or translate them into multiple different terms in the target language. The translation will remain hidden to the end user. This approach of using "rough" translations for retrieval is both cheaper to implement and often more effective for retrieval, as multiple translation alternatives can be retained during the process.

4.2 Outline of Approach, Experiment Setup and Retrieval

The outline of this approach, called in the following "pseudo-translation of expanded records", or "pseudo-translation" for short, is thus: (1) the unstructured content fields of the record are expanded by additional terms that are statistically similar to the original terms in those fields; (2) these additional terms are derived by searching for records that are similar in content to the record that is to be expanded, and then extracting from these additional records the best terms according to well-known blind feedback techniques; (3) the expanded records are translated in their entirety using a simple translation resource; and (4) retrieval takes place on the expanded, (pseudo-) translated records.

With retrieval experiments on test collections, we have aimed to demonstrate how expanded records could be represented, how they would look in their pseudo-translated state and to analyze whether they could be expected to be usable for implementing CLIR in the TEL system.

A full evaluation on a sample of 151,700 bibliographical records in English from the British library (part of the TEL central index) was carried out. We used 99 queries in English derived from three months of logfiles to represent typical information needs. Queries are a mix of one-word statements and longer formulations. The queries were manually translated into German for later cross-language retrieval experiments.

The experiments follow the so-called Cranfield paradigm [9] for retrieval tests. The retrieval system used was Terrier[4], an open-source information retrieval system developed by the University of Glasgow. Note that much of the procedures described would be implemented off-line in an operational system. Translation of the records was done using the PROMT[5] off-the-shelf machine translation system. Again, a variety of different translation resources could be used in support for the chosen CLIR approach.

We expanded the 151,700 records by using each record in turn as a query and running it against the whole collection to determine the set of most similar records. The 10 best-ranked items were used to produce a maximum of 5 expansion terms leading to the most promising results. For some records, no new statistically associated terms can be found, and the records remain unexpanded. In all, approximately 29% of records were not expanded. This ratio should drop if more records were added to the "document" base.

We pseudo-translated all expanded records from English to German using PROMT. The translation suffered from aggressive compound formation by the PROMT software. Since we did not have a German compound splitter available for the Terrier system, retrieval effectiveness may have been negatively affected (For the effect of "decompounding" on retrieval effectiveness, see e.g. [8]).

The following is an example of a pseudo-translated record from our test collection:

English record, original:

```
<srw_dc:dc><recordPosition>103899</recordPosition>
<title>Private power : Multinational corporations for the
survival our planet.</title></srw_dc:dc>
```

German record, pseudo-translated, expanded.

```
<srw_dc:dc><recordPosition>103899</recordPosition>
<title>Private Macht : Multinationale Vereinigungen für das
Überleben unser Planet.</title>
<extendedTerms>Entwicklung, Welt, </extendedTerms></srw_dc:dc>
```

The resulting 151,700 pseudo-translated records were loaded into the Terrier system for retrieval.

The 99 queries were hand-translated into German and used to retrieve the top 10 records for each query from the pseudo-translated German records. This constitutes a cross-language retrieval experiment, as each pseudo-translated record can clearly be matched with the original English version it represents in the search index. As a

[4] Terrier is available under the Mozillla Public License.

[5] http://www.e-promt.com/

baseline for comparison, we ran the same 99 queries in their original English version against the original English records.

4.3 Analysis of Results

To evaluate retrieval effectiveness, usually recall and precision figures are calculated. Clearly, it was not feasible to do extensive manual relevance assessments for all 99 queries in our study (resulting in 151,700 * 99 assessments). We used so-called "overlap analysis" as a viable alternative. The monolingual English baseline, representing the same information need as the cross-language case, acts as a "gold standard", by assuming that the results from that retrieval experiment have sufficient quality. Any retrieval result sufficiently similar to the monolingual result is then considered to be acceptable. We analyzed the top 10 ranked records to determine the similarity between the monolingual and the cross-language experiment. In all, 30 of the 99 queries had sufficiently similar results, and thus the cross-language results were considered to match the monolingual baseline. These queries were excluded from further analysis.

The remaining 69 queries have results that significantly differ from the monolingual baseline. This, however, does not necessarily indicate that these queries have poor performance. For further analysis, four cases need to be distinguished: (1) good monolingual result; good, but different, cross-language result; (2) good monolingual result; bad cross-language result; (3) bad monolingual result; good cross-language result; (4) bad monolingual result; bad, but different, cross-language result. We supplement this with the previous case: (0) monolingual and cross-language result similar; assumed to be good.

We attempted to classify the remaining 70 queries (one query was accidentially duplicated at this stage) to cases 1-4 based on relevance assessments of the top 10 records for both the monolingual and cross-language experiments. In combination with the actual analysis of the results, it was not possible to process all remaining queries. A total of 18 queries had to be excluded from further processing due to lack of resources. We thus analyzed a grand total of 52 queries, giving a categorization for 82 queries.

We argue that case 0, 1, and 3 provide evidence for good retrieval results, whereas case 4 at least indicates that the cross-language result is not necessarily worse than the monolingual result. In all, using this methodology we found that 55% of queries analyzed showed evidence of good retrieval results, and 83% of queries showed evidence that they did not suffer significantly from the cross-language setup when compared to the monolingual baseline (note that for some of these queries there simply will be no relevant records in the collection!). The latter number is encouraging, being in line with what has been reported as state-of-the-art for CLIR in the CLEF campaign for lengthy documents [4]. Please note, however, that the number has to be treated with care, owing to the limitations described above. The approach should actually benefit in terms of effectiveness when scaling up to larger collections.

Table 1. Summary of evaluation of queries

Case	0	1	2	3	4	not eval.
# queries	30	13	14	2	23	18

5 Conclusions

We have described the results and the findings of a feasibility study carried out to determine how multilingual information access functionalities could be added to the TEL system. We have proposed two different approaches for introducing MLIA functionalities in the TEL system: the first one, called "isolated query translation", performs a pre-processing step to translate the query and then routes the translated query to the national library systems. The second one, called "pseudo-translation", involves only queries sent to the TEL central index but merges the translation process with the retrieval one in order to offer more effective MLIA functionalities. Please note that the two approaches are independent, and we expect considerable potential for combination.

On the whole, we can envision the following evolutionary scenario for implementing MLIA in TEL:

- short-term: the "isolated query translation" solution is a first step for adding MLIA functionalities to TEL and represents a quick way to give TEL users and partners a multilingual experience.
- mid-term: the implementation and deployment of a "pseudo-translation" solution is a second step which better exploit the information directly managed by the TEL central index;
- long-term: the adoption of an inter-lingua approach, where all the translations are made to and from this pivot language, will allow for scaling up the system, when new partners will join TEL. This can be facilitated by combining the two approaches described in this paper.

The work for defining an actual roadmap for introducing MLIA functionalities into TEL is currently ongoing and some initial results in this direction are reported in [1].

Acknowledgments

We thank Marco Dussin for his work on TEL user interface. Many thanks are also due to Bill Oldroyd of the British Library for his assistance in obtaining the set of records used for the experiments on pseudo-translated, expanded records. Thanks also go to Eric van der Meulen and Julie Verleyen of the TEL office for their help with the current TEL architecture. Thomas Arni helped with running the pseudo-translation experiments, and Carol Peters provided corrections to the paper.

The work reported in this paper has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

References

[1] Agosti, M., Braschler, M., Ferro, N., Peters, C., Siebinga, S.: Roadmap for MultiLingual Information Access in The European Library. In: ECDL 2007. Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries. LNCS, vol. 4675, pp. 136–147. Springer, Heidelberg (2007)

[2] Ballesteros, L., Croft, W.B.: Phrasal translation and query expansion techniques for cross-language information retrieval. In: Proc. 20th Annual International ACM SIGIR Conference, pp. 84–91. ACM Press, New York (1997)
[3] Braschler, M., Peters, C.: Cross-Language Evaluation Forum: Objectives, Results, Achievements. Information Retrieval 7(1/2), 7–31 (2004)
[4] Braschler, M.: Robust Multilingual Information Retrieval Dissertation. Institut Interfacultaire d'Informatique, Université de Neuchâtel (2004)
[5] Braschler, M., Di Nunzio, G.M., Ferro, N., Peters, C.: CLEF 2004: Ad Hoc Track Overview and Results Analysis. In: Peters, C., Clough, P.D., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (eds.) CLEF 2004. LNCS, vol. 3491, pp. 10–26. Springer, Heidelberg (2005)
[6] Braschler, M., Ferro, N., Verleyen, J.: Implementing MLIA in an existing DL system. In: MLIA 2006. Proc. International Workshop on New Directions in Multilingual Information Access, pp. 73–76 (2006)
[7] Braschler, M.: Combination Approaches for Multilingual Text Retrieval. Information Retrieval 7(1/2), 183–204 (2004)
[8] Braschler, M., Ripplinger, B.: How Effective is Stemming and Decompounding for German Text Retrieval? Information Retrieval 7(3/4), 291–306 (2004)
[9] Cleverdon, C.W.: The Cranfield tests on index language devices. Aslib Proceedings 19, 173–192 (1967), Reprinted in (Sparck Jones and Willett, 1997)
[10] Di Nunzio, G.M., Ferro, N., Jones, G.J.F., Peters, C.: CLEF 2005: Ad Hoc Track Overview. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 11–36. Springer, Heidelberg (2006)
[11] Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C.: CLEF 2006: Ad Hoc Track Overview. In: CLEF 2006. Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum. LNCS, vol. 4730, pp. 21–34. Springer, Heidelberg (2007)
[12] Ferro, N., Braschler, M., Arni, T., Peters, C.: Deliverable D8.3.1 – Feasibility study on Multilinguality. DELOS, A Network of Excellence on Digital Libraries – IST-2002-2.3.1.12, Technology-enhanced Learning and Access to Cultural Heritage (2006)
[13] Frakes, W.B., Baeza-Yates, R.: Information Retrieval. In: Data Structures & Algorithms, Prentice-Hall, Englewood Cliffs (1992)
[14] Peters, C., Braschler, M.: European Research Letter: Cross-language system evaluation. The CLEF campaigns, Journal of the American Society for Information Science and Technology 52(12), 1067–1072 (2001)
[15] van Veen, T., Oldroyd, B.: Search and Retrieval in The European Library. A New Approach. D-Lib Magazine 10(2) (2004)