CACAO System: An Overview

Raffaella Bernardi[2], Massimo Balestrieri[1], Alessio Bosca[3], Luca Dini[3], Daniele Gobbetti[2], and Frédérique Segond[4]

[1] Gonetwork s.r.l.,
[2] Faculty of Computer Science, Free University of Bozen-Bolzano,
[3] CELI, s.r.l.,
[4] Xerox Research Centre Europe

Abstract. This extended abstract gives an overview of the CACAO system highlighting its innovative features, its current status and its next expected steps. CACAO is a two year project funded by the eContent*plus* Program carried out in collaboration by experts in Language Technologies and Digital Libraries.

1 Background

The large amount of on-line catalogues and digitalized documents across Europe launches two main challenges tightly connected to each other. On the one hand, there is the need of allowing users to search through the different catalogues simultaneously and on the other hand to find books on relevant topics even if in different languages. Language Technologies (LT) help achieving both tasks. An interesting example of how they can improve searchability in aggregated collections is provided in [5]; the system we present in this paper, CACAO[5] (Cross-language Access to Catalogues And On-line libraries), is mainly an answer to the multilingual retrieval challenge, though it necessarily concerns the aggregation problem too. Both tasks require a close collaboration between LT and (Digital) Libraries experts. CACAO is the result of joint efforts by experts of both fields.

CACAO project proposes an innovative approach for accessing, understanding and navigating multilingual textual content in digital libraries and library catalogues. By coupling sound Natural Language Processing techniques with available information retrieval systems the project aims at the delivery of a non-intrusive infrastructure to be integrated with current OPAC and digital libraries. The result of such integration will be the possibility for the user to type in queries in his/her own language and retrieve volumes and documents in any available language.

The system has been already put at work over the catalogues of the five consortium libraries[6] and tested on The European Library catalogues.

[5] CACAO is an EU project supported by the eContent*plus* Programme of the European Commission (ECP 2006 DILI 510035): `http://www.cacaoproject.eu`

[6] The Göttingen University Library (German), Library of the Free University of Bozen-Bolzano (Italian, English and German), Cité des Sciences et de l'Industrie (French), Kornik Library (Polish), National Széchényi Library (Hungarian).

The main objective of CACAO was crossing the chasm between sound innovation and adoption by library institutions for real life purposes. The test cases mentioned above show the success of the undertaken approach and the achievement of CACAO main goal. The system can now be further improved and fine-tuned thanks to the built infrastructure, its application to several different catalogues, and the collection of end-users queries' logs that we will be able to obtain through the time.

2 The CACAO solution

The architecture of the CACAO system, summarized in Figure 1, is an integration of several subsystems coordinated by a central manager that triggers scheduled activities (i.e. data harvesting or processing) and reacts to external stimuli represented by end users queries. The "Harvesting" subsystem is in charge of collecting data from digital libraries, abstracting from the multiplicity of standards and protocols, and storing them in a repository. The "Corpus Analysis" subsystem performs specific analysis and transformations on the data collected from libraries and infers new information that is then used to support query processing and resource retrieval (e.g. query expansion, terms disambiguation). The "CLIR" (Cross Language Information Retrieval) subsystem is in charge of analyzing the monolingual user query in input and transforming and enriching it by means of translations and expansions. Finally, the "Web Services" subsystem represents external modules providing specific services (e.g. linguistic analysis, translations).

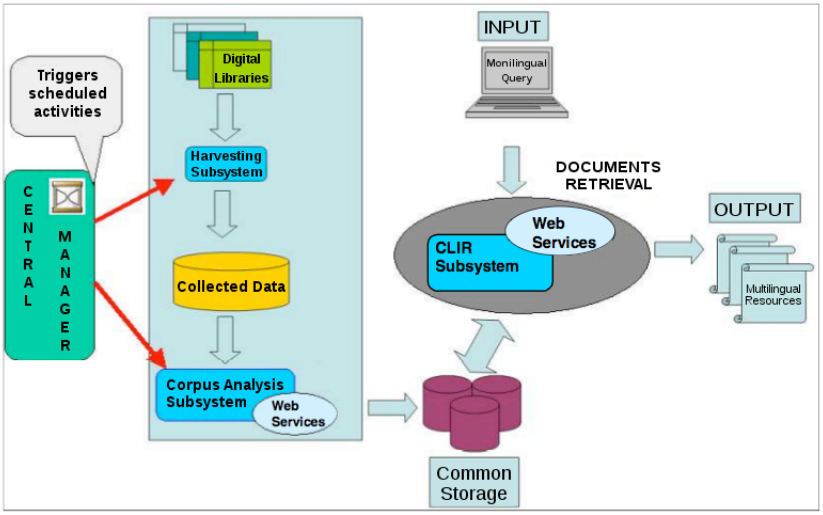

Fig. 1. CACAO architecture

The CLIR subsystem currently handles the languages of the consortium libraries, viz. German, Italian, English, French, Polish and Hungarian, and its flexible architecture allows easy integration of new resources in order to support new languages too.

In order to monitor its development CACAO has been evaluated already in its early stage via the participation to the TEL@CLEF campaign in 2008 [2]; the dataset of this campaign has been further used to evaluate the role of the different modules and resources that were developed during the project period. In particular, it has been used to enhance the system with the TLike algorithm, a translation algorithm that, by identifying in users logs whether a query is a likely translation of a previously submitted one, enriches CACAO translation resources. (See [3] for details.) A second module which is under evaluation for the different CACAO languages is the "Word To Category" (W2C) module. It has been described in [1] where a first evaluation in a controlled experiment has been reported for German, Italian and English.

CACAO can either be installed locally to obtain a Cross Language Access to the Catalogue of a certain library, or used in a federated setting. In both cases, users can access it via a Google like simple interface provided by the consortium and easy customizable, or via a facet-based advanced interface. Both interfaces will be soon available via the project web site `http://www.cacaoproject.eu`. Currently, the consortium is completing the harvest of TEL records and constructing three Thematic Portals on European History, Geography and Mathematics. To facilitate these aggregations, CACAO has developed an Application Profile based on The European Library Application Profile for Objects, discussed in [4].

3 Conclusion

CACAO experience shows that LT are in a mature stage for being applied to real life tasks, as the one required by the Digital Libraries world; furthermore such tasks and the possibility to analyse real users' behaviours and requests launch interesting new challenges and increase the appeal of these research field. CACAO is already running over five libraries catalogues covering six European languages, and is currently harvesting data from TEL. Therefore, it will soon become a useful source for log analysis and any further evaluation of LT related modules that can be easily integrated into its architecture. The consortium is now at work on collecting end-users feedback, as well as evaluating the TLike algorithm and the W2C module that are expected to further improve CACAO precision.

References

1. R. Bernardi, D. Gobbetti, and L. Siciliano. Multilingual access to library catalogues: Word sense disambiguation via classification systems. In *Proceedings of the International Conference on Semantic Web and Digital Libraries*, 2009. In Printing.

2. A. Bosca and L. Dini. Query expansion via library classification systems. In *CLEF@TEL*, 2008.
3. A. Bosca and L. Dini. The role of logs in improving cross language access in digital libaries. In *Proceedings of the International Conference on Semantic Web and Digital Libraries*, 2009. In Printing.
4. B. Levergood, L. Siciliano, and S. Chambers. An application profile for interoperable and reusable metadata in a cross-language context. In R. Bernardi, S. Chambers, and B. Gottfried, editors, *Proceedings of The Workshop on Advanced Technologies for Digital Libraries (AT4DL 2009)*. University Library of Bozen/Bolzano, 2009. This volume.
5. D. Newman, K. Hagedorn, C. Chemudugunta, and P. Smyth. Subject metadata enrichment using statistical topic models. In *JCDL*, 2007.