CACAO PROJECT AT THE TEL@CLEF 2009 TASK

Alessio Bosca, Luca Dini
Celi s.r.l. - 10131 Torino - C. Moncalieri, 21
alessio.bosca, dini@celi.it

Abstract

This paper presents the participation of the CACAO prototype to the TEL@CLEF 2009 task, an evaluation track focusing on multilingual document retrieval over a collection of library catalogues. CACAO (Cross-language Access to Catalogues And On-line libraries) is an EU project devoted to enabling cross-language access to the contents of a federation of digital libraries with a set of software tools for harvesting, indexing and serching over such data. CACAO project consortium participated both to the monolingual and the bilingual subtasks from TEL@CLEF since they constitute a perfect opportunity in order to test the CACAO system prototype and obtain feedbacks for its enhancement. The prototype showed good performances with respect to the other participants, resulting among the best 5 in the French Monolingual subtrack, and quite a significative enhancement in comparison with the past year participation.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-Language Information Retrieval, Query Expansion, Translations Disambiguation, Digital Libraries

1 Introduction

The TEL@CLEF is an evaluation track proposed to the CLEF campaign participants with the aim of searching and retrieving relevant items from collections of library catalogues; the bibliographic metadata is divided into 3 collections, respectively extracted from British, French and Austrian national libraries.

TEL@CLEF track offers a set of subtasks reflecting the multilinguality of the data, respectively focusing on monolingual and bilingual information retrieval; 50 topics has been prepared for each of the 3 main collection languages and each topic has 2 fields: a title with 2-4 key terms and a description field, containing a sentence that specifies in more detail the information needs of the user.

CACAO project consortium participated both to the monolingual and the bilingual subtasks from TEL@CLEF since they constitute a perfect opportunity in order to test the CACAO system

prototype and obtain feedbacks for its enhancement. The prototype showed good performances with respect to the other participants, resulting among the best 5 in the French Monolingual subtrack, and quite a significative enhancement in comparison with the past year participation.

This paper is organized as follows. We present the architecture of our system in 2, in 3 we describe our experiments, the evaluation measures and the evaluation results, and finally conclude in 4.

2 CACAO Project

CACAO (Cross-language Access to Catalogues And On-line libraries) is an EU project funded under the eContentplus program and proposes an innovative approach for accessing, understanding and navigating multilingual textual content in digital libraries and OPACs, enabling European users to better exploit the available European electronic content.

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. CACAO aims at offering cross-lingual and cross-border access to the content of classical and digital libraries and enabling users to find digital content irrespective of the language. In fact, in a context of interlaced cross-border libraries, such as the ones proposed by META OPAC, the absence of a cross-language perspective is likely to cause a substantial impasse: if a user wanted to access a META OPAC including the National Libraries of France, Germany, Italy, Poland and Hungary, s/he would have to type five queries in five different languages. Much of the advantage of having a unique access point is thus lost.

CACAO project proposes a system based on the assumptions that users look more and more at library contents using free keyword queries (as those used with a web search engine) rather than more traditional library-oriented access (e.g. via Subject Heading); therefore, the only way to face the cross-language issue is by translating the query into all languages covered by the library/collection (rather than, for instance, translating subject headings, as in the MACS approach, https://macs.vub.ac.be/pub/). The system will then yield results in all desired languages.

Validation is another important aspect in the project: all CACAO core technologies are indeed sound, but they have never been massively deployed in the field of digital libraries. CACAO aims at crossing the chasm between sound innovation and adoption by library institutions for real life purposes. CACAO proposes the development of an infrastructure for multilingual access to digital content, including an information retrieval system able to search for books and texts in all the available languages. The core of the search engine takes advantage of information contained in existing catalogues and texts of the digital libraries that is enriched by means of NLP techniques such as word sense disambiguation and named entities recognition. The goal of such integration is to avoid confusing the user by providing irrelevant results due to bad translations and thus enabling a better access to the digital content.

2.1 Architecture Overview

The general architecture of the Cacao system could be summarized as the result of the interactions of few functional subsystems, coordinated by a central manager and reacting to external stimuli represented by end users queries:

- Harvesting subsystem is in charge of collecting data from digital libraries, abstracting from the multiplicity of standards and protocols, and storing them into a repository.
- Corpus Analysis subsystem performs specific analysis on the data collected from libraries and infers new information used to support query processing and resource retrieval (e.g. query expansion, terms disambiguation,..).

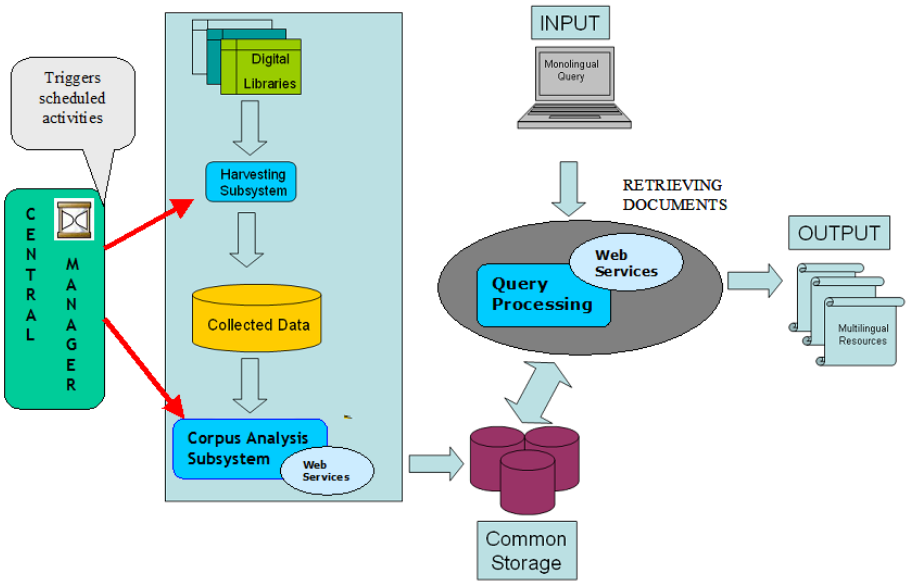

Figure 1: CACAO System Architecture

- Web Services subsystem represents third party software providing specific services (e.g. linguistic analysis, translations,..).
- Query Processing subsystem: a set of components is devoted to process the original monolingual user query, transforming and enriching it by means of translations and expansions.

3 Experiments

3.1 Indexing and Enhancing MetaData

In order to acquire the collections metadata into the CACAO system a specific harvester module for importing the XML corpus documents has been deployed. The textual information contained in the *dc:subject*, *dc:title* and *dc:dexcription* are lemmatised using the XIP incremental parser from XEROX (see [1])) and all the data is then indexed using the Lucene open source engine (see [3])). By means of lexical semantics technologies (we exploited Random Indexing approach, see [2]) a corpus based word space model has been created for each of the TEL@CLEF collections; these word space resources have been used by the CACAO system as a means to disambiguate the candidate translations and for query expansion purposes.

3.2 Topics Processing

The approach adopted by CACAO system for dealing with user queries is based on the free keywords search; therefore while the title field of TEL topics already fitted this model, the description field has been processed in order to extract a set of relevant keywords from the sentence. For this purpose a simple keyword extractor module has been used for each of the main languages present in the corpus (English, French and German).

The keywords retrieved in this process are lemmatised and the system assigns different weight to them according to their frequency in both of the topic fields (title and description). In the lemmatisation process named entities (i.e. person and geographic names) were also identified, and they were treated differently from common keywords with respect to translation to target languages.

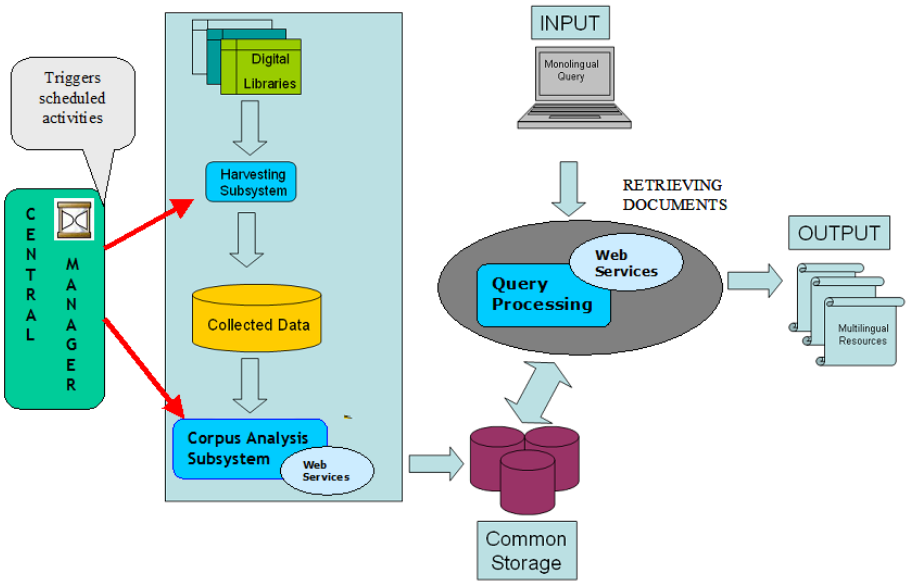

Figure 2: Topic Processing

According to the subtasks (monolingual or bilingual) the keywords were translated to the target language or directly submitted to the Lucene search engine.

The translation process exploits internal resources (inter-lingual indexes or bilingual dictionaries) and online dictionaries as Ergane; the obtained translation candidates are disambiguated using the corpus based semantic vectors, computed by the CACAO system on the collections metadata 2.1 and according the following approach:

- As a first step the system automatically groups keywords in sets of semantically related terms by comparing their similarity, defined as the cosine of the angle between the vector representations of the terms; this process allows the system to group together all the keywords bearing a common meaning.
- Then the translation candidates of each keywords group are analysed in order to prune away all the elements with a low similarity with the center of the translation group, computed as the sum of the vector representation of terms (a variation of the algorithm proposed by [4]).

Experiments involving query expansion enriched the keywords groups (either in the original or in the target language) by means of different strategies :

- Exploiting the corpus based semantic vectors, by adding the N nearest neighbours of each group center where; the actual value of N depends on the cardinality of the keyword group.
- Extracting terms from the titles of the top 10 documents restrieved for each topic
- By expanding geographic names items with the list of the geographic entity they are contained in (i.e. Turin is expanded as Piedmont, Italy, Europe)
- By translating the terms in the title field of TEL topics to other languases then the default one of the collection in order to capture the multilinguality of the data (i.e. in the monolingual English task, keywords from the title filed are also translated to french and german)

3.3 Submitted Runs

The results of the top result for each subtask are provided in the following table:

SubTask	Run ID	MAP	R-Prec
En Monolingual	SEMVECT_EXPANDED	30,54%	26.08%
En Bilingual	DE-EN_BASE	17.10%	19.23%
Fr Monolingual	ML_EXPANDED	27.35%	23.61%
Fr Bilingual	EN-FR_ML_EXPANDED	14.23%	16.45%
De Monolingual	ALL_EXPANDED	21.02%	24.23%
De Bilingual	EN-DE_BASE	9.05%	11.22%

Table 1: Submitted Experiments

4 Conclusions

The prototype showed good performances with respect to the other participants, resulting among the best 5 in the French Monolingual subtrack, and quite a significative enhancement in comparison with the past year participation.

5 Acknowledgements

This work has been supported and founded by CACAO EU project (ECP 2006 DILI 510035).

References

[1] At-Mokhtar S., Chanod J-P., Roux C. Robustness beyond shallowness: incremental dependency parsing NLE Journal, 2002.

[2] Sahlgren, M. An Introduction to Random Indexing. Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, August 16, Copenhagen, Denmark.

[3] Lucene. The Lucene search engine. URL: http://jakarta.apache.org/lucene/.

[4] A. Bosca and L. Dini. Query expansion via library classication systems. LNCS proceedings on CLEF@TEL, 2008.