CACAO PROJECT AT THE TEL@CLEF 2008 TASK

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

Abstract: The paper describes the participation of the CACAO project consortium to the TEL@CLEF 2008 task targeted at retrieving relevant items from collections of library catalogues. 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. For each monolingual and bilingual subtask two different experiments have been conducted, one involving additional query expansion and one not. Results evidenced a poor initial performance of the system, however since the project started few months ago they can constitute a valuable baseline in order to measure the future advancement of the system.

ACM Categories: H.3.3 Information Search and Retrieval, H.3.7 Digital Libraries

Keywords: cross-language information retrieval, query expansion, translations disambiguation, bibliographic data, digital libraries.

1 Introduction

The TEL@CLEF is a new task proposed this year to the CLEF campaign participants with the aim of searching and retrieving relevant items from collections of library catalogues. The data is different from the corpora previously used in the CLEF ad hoc tracks and consists of bibliographic metadata divided into 3 collections, extracted from British, French and Austrian national libraries. TEL@CLEF task 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.

In our participation to the TEL@CLEF task we focused both on the monolingual and on the bilingual retrieval subtasks since CACAO project aims at offering a cross-language acces to the contents of a federation of digital libraries and the tasks constitute a perfect opportunity to test the baseline version of the CACAO system prototype and obtain feedbacks for its enhancement.

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

2.1 Architecture Overview

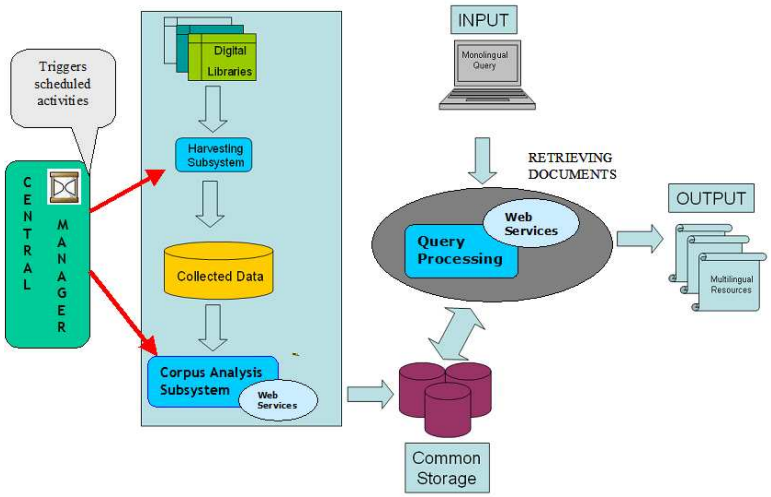

Figure 1 - CACAO architecture

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.

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,..).
- **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 Description

For each monolingual and bilingual subtask two different experiments have been conducted, one involving additional query expansion and one not; in the following subsections the experiments set-up, topic processing and experimental results are described.

3.1 Experiments Set-up

In order to acquire the metadata of TEL@CLEF collections into the CACAO system, GoNetwork s.r.l. (one of the CACAO partners) deployed a specific harvester module for importing the XML corpus documents. The

textual information contained in the *dc:subject*, *dc:title* and *dc:dexcription* has been lemmatised using the XIP incremental parser from XEROX (see [1]) and all the data has been then indexed using the Lucene open source engine (see [4]).

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).

Each description sentence has been analysed in order to extract two different kinds of information, one representing the content type of the items to be retrieved (as novels, poetry or photo collections) and the other conveying additional detail on user interests.

The keywords retrieved in this process have been lemmatised and the system assigned different weight to them according to their frequency in both of the topic fields (title and description). In the lemmatisation process named entities were also identified, as they have been treated differently from common keywords with respect to translation to target languages. 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 exploited internal resources (inter-lingual indexes or bilingual dictionaries) and online dictionaries as Ergane (see [3]); the so-obtained translation candidates have been disambiguated using the corpus based semantic vectors, computed by the CACAO system on the collections metadata (see subsection 3.1) 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 [5]).

Experiments involving query expansion enriched the keywords groups (either in the original or in the target language) exploiting the corpus based semantic vectors by adding the *N* nearest neighbours of each group center (not already present in the keyword set), where the actual value of *N* depended on the cardinality of the keyword group.

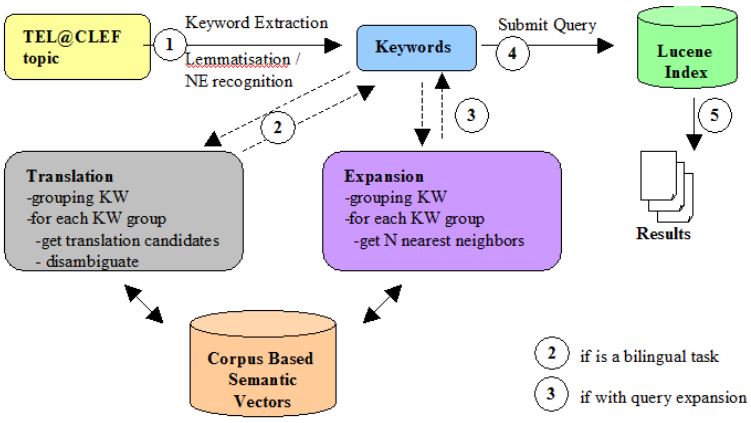

Figure 2 - Topic Processing

3.3 Submitted Runs and Evaluation Results

We submitted two runs for each monolingual and bilingual subtask for each target language for a total of 18 runs. The results of these experiments are provided in the following tables where results are separated for target language.

Run ID	Source Language	MAP	Average Precision	Relevant Docs	Relevant Docs retrieved
CacaoEngEngPlain	en	17,27%	0,173	50,66	32,5
CacaoEngEngExpanded	en	13,30%	0,133	50,66	28,82
CacaoFreEngPlain	fr	5,75%	0,058	50,66	13,98
CacaoFreEngExpanded	fr	4,35%	0.044	50,66	12,6
CacaoGerEngPlain	de	4,89%	0,049	50,66	12,54
CacaoGerEngExpanded	de	4,61%	0,046	50,66	11,78

Table 1- Target Language: English

Run ID	Source Language	MAP	Average Precision	Relevant Docs	Relevant Docs retrieved
CacaoFreFrePlain	fr	16,97%	0,17	26,78	14,84
CacaoFreFreExpanded	fr	12,98%	0,13	26,78	13,4
CacaoEngFrePlain	en	6,78%	0,068	26,78	9,02
CacaoEngFreExpanded	en	5,51%	0,055	26,78	8,24
CacaoGerFrePlain	de	1,93%	0,019	26,78	1,72
CacaoGerFreExpanded	de	1,53%	0,015	26,78	1,64

Table 2- Target Language: French

Run ID	Source Language	MAP	Average Precision	Relevant Docs	Relevant Docs retrieved
CacaoGerGerPlain	de	11,46%	0,115	32,74	16,39
CacaoGerGerExpanded	de	9,42%	0,096	32,74	15,56
CacaoEngGerPlain	en	4,25%	0,043	32,74	9,78
CacaoEngGerExpanded	en	3,43%	0,034	32,74	8,68
CacaoFreGerPlain	fr	4,04%	0,04	32,74	6,5
CacaoFreGerExpanded	fr	3,18%	0,032	32,74	5,94

Table 3- Target Language: English

4 Conclusion

CACAO project has started only few months ago and despite the early stage of development of the system prototype the consortium decided to participate to the TEL task of the CLEF campaign in order to asses a baseline of the system and hence measure the advancement of the system during the project timeline.

Compared to other participants the experimental results evidenced a poor initial performance of the CACAO system; however such poor performance is partly due to a misinterpretation of the task: we in fact assumed that the task was *completely* partitioned according to languages. Therefore when performing a bilingual subtask (i.e. the task DE->FR on the TEL data of BNF, the French National Library) we assumed that only the elements in the target language should be retrieved, while the golden standard over which evaluation was performed actually

contains a not negligible percentage of results in other languages. All these records were not even searched, dramatically lowering recall and other measures.

5 Acknowledgments

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

6 References

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

[2] Sahlgren, M. (2005): 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] Ergane. An online multilingual dictionary. URL: http://download.travlang.com/Ergane/

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

[5] Curtoni P. and Dini L., Celi participation at CLEF 2006: Cross language delegated search. In CLEF2006 Working notes