CACAO PROJECT AT THE LOGCLEF TRACK

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 Log Analysis for Digital Societies (LADS) task of LogCLEF 2009 track. 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. In our experiment we investigated the possibility to exploit the TEL logs data as a source for inferring new translations, thus enriching already existing translation dictionaries; the proposed approach is based on the assumption that users consulting a multilingual digital collection are likely to repeat the same query in different languages. We applyed our approach to the logs from TEL and the results obtained are very promising.

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.3.7 Digital Libraries

General Terms

Measurement, Performance, Experimentation

Keywords

Cross-Language Information Retrieval, Log Analysis, Translations Disambiguation, Digital Libraries

1 Introduction

The Log Analysis for Digital Society (LADS) from LogCLEF track is a new task that focuses on the log analysis as a means to infer new knowledge from user logs (i.e. users behaviours, multilingual resources); in particular the task proposes to the participants to deal with logs from The European Library (TEL).

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.

In our experiment we focused on the multilinguality aspect of log analysis and in particular we investigated the possibility to exploit the TEL logs data as a source for inferring new translations and thus enriching already existing translation resources for dictionary based cross language access to digital libraries.

The proposed methodology is based on the assumption that when users are aware of consulting a multilingual digital collection, they are likely to repeat the same query several times, in several

languages. By adopting the proposed algorithm, it is possible to discover translationally equivalent queries in logs produced by monitoring user queries.

The basic idea beyond our approach (named TLike algorithm) is to detect the probability for two queries to be one a translation of the other. In the simple case we expect that if all the words in query QS have a translation in query QT and if QS and QT have the same number of terms, then QS and QT are translation equivalent. Things are of course more complex than this, due to the following facts:

- The presence of compound words make the constraints on cardinality of search terms defeasible (e.g. the Italian carta di credito vs. the German KreditCarte).
- One or more words in QS could be absent from translation dictionaries.
- One or more words in QS could be present in the translation dictionaries, but contextually correct translation might be missing.
- There might be items which do not need to be translated, notably Named Entities.

This paper is organized as follows. We present the architecture of our system in 2, in 3 we describe our experiments and the obtained results; we 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.

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,..).

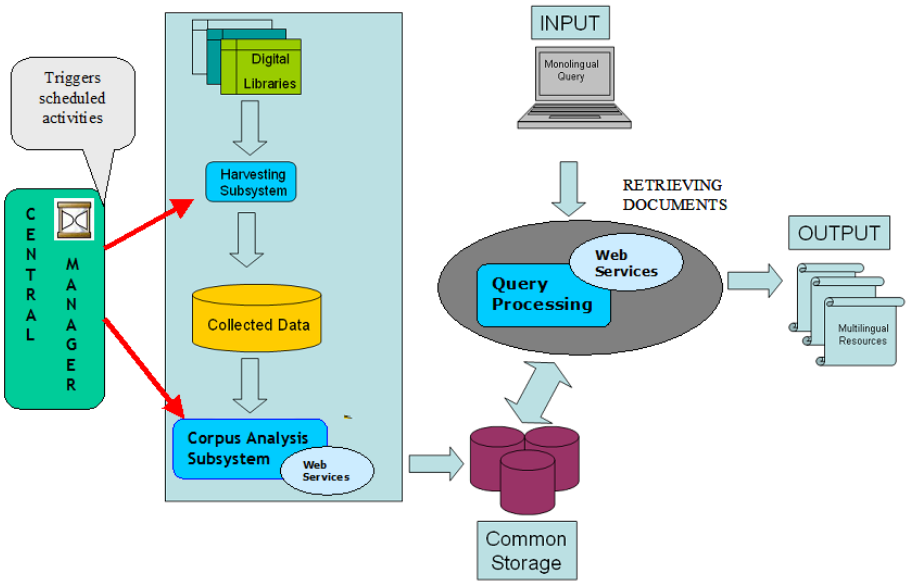

Figure 1: CACAO System Architecture

- 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

The first step of our experiment consisted in the creation of a Lucene Search index starting from the TEL logs; information contained in the query field of the logs has been filtered in order to remove terms pertaining to the query syntax (restrictions on fields, boolean operators,...) enriched by means of shallow NLP techniques as lemmatization and named entities recognition and of a language guesser facility used to individuate the query source language.

The second step involved the CACAO search engine in order to create a resource containing all possible translation candidates. The CACAO project (see [2]) aims at providing a European level answer to cross language information access to digital libraries, by exploiting the by now mature technology of cross language information retrieval.

Each distinct query contained in the logs has been used as input for the CACAO system in order to obtain a set of translation candidates for the query. In fact CACAO system translated the query from the TEL logs into all the languages natively supported (english, french, german, polish, hungarian and italian) and then exploited such translations in order to search for related queries in other languages; the result of this step consisted in a textual file containing for each distinct query a list of translation candidates proposed by the CACAO engine.

The third step of the algorithm consisted in exploiting the T-Like procedure in order to evaluate the probabilities associated to the different translations candidates extracted from the logs and thus obtain a list of proposed translations as well as some statistics on the retrieved translations.

The TLike algorithm is based on three main resources:

- A system for Natural Language Processing able to perform for each relevant language basic tasks such as part of speech disambiguation, lemmatization and named entity recognition.
- A set of word based bilingual translation modules.

- A semantic component able to associate a semantic vectorial representation to words.

The basic idea beyond the TLike algorithm is to detect the probability for two queries to be one a translation of the other and a detailed description of the strategy adopted can be found in [3].

3.1 Experiment results

Table 1 presents an excerpt of the translation pairs extracted from the TEL logs with our approach while table 2 shows some statistic measure on the retrieved translations.

Source Query in Logs	Candidate Translations from Logs
the road to glory [en]	en route pour la gloire [fr]
la vita di gesu narratasales [it]	essai sur la vie de jsus [fr]
die russische sprache der gegenwart [de]	russian language composition and exercises [en]
democratie [fr]	the future of democracy [en]
digital image processing [en]	cours de traitement numrique de l image [fr]
biblia krolowej zofii [pl]	simbolis in the bible [en]
architecture [en]	trattato di architettura [it]
inondation [fr]	after the flood [en]
guerre mondiale [fr]	guerra mondiale [it]
quali varieta di meli e di peri [it]	biology of apple and pear [en]
national library of norway [en]	biblioteka narodowa [pl]
portrait de dorian gray [fr]	the portrait of dorian gray [en]
la guerre et la paix [fr]	war+and+peace [en]
production de l espace [fr]	the production of space [en]
exposition universelle 1900 [fr]	esposizione universale di roma [it]
storia della chiesa [it]	church history [en]
firmen landwirtschaftliche maschinen [de]	lagriculture et les machines agricoles [fr]
lord of the rings [en]	le seigneur des anneaux [fr]
dictionnaire biographique [fr]	dizionario biografico [it]
deutsche mythologie [de]	the mythology of aryan nations [en]
ancient maps [en]	carte antique [fr]
round the world in 80 [en]	le tour du monde en 80 [fr]

Table 1: Submitted Experiments

true Positive translations	351.0
true Negative translations	0.0
false Positive translations	0.0
false Negative translations	80049.0
Precision	1.0
Recall	0.004365671641791045

Table 2: Evaluation Measures

4 Conclusions

This paper represents the first step of a research on NLP based query log analysis. The preliminary results are quite incouraging and in the future we plan to extend this research into two directions:

- We will consider all the information contained in query logs, such as session identifiers, temporal distance, repetition of the same query, semantic distance among similar queries, etc.
- We will try to extend the semantic matching method to cover cases where the semantic vectors are not present in the semantic repository. This will imply the use of the web and web search engines as a dynamic corpus([4]).

5 Acknowledgements

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

References

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

[2] A. Bosca and L. Dini. Query expansion via library classication systems. LNCS proceedings on 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.

[4] Baroni, M., Bisi, S.: Using cooccurrence statistics and the web to discover synonyms in technical language