CELI participation at ImageCLEF 2006: Comparison with the Ad-hoc track

Paolo Curtoni
Luca Dini
Vittorio Di Tomaso
CELI
Torino-Italy
curtoni/dini/ditomaso@celi.it

Abstract

In this paper we discuss the CELI's first year of activity at CLEF in the context of Cross Language Image Retrieval Track (ImageCLEF). The proposed system is an upgrade of CELI's cross language delegated search system (www.elois.biz). The system is meant to perform CLIR on the web by using Google and Yahoo indexes. Therefore the goal is to provide reasonable translations of queries with no direct access to the corpus. This, in turn, means absence of tuning procedure for the system and impossibility to impose restrictions in terms of domain, style, etc. Our approach is based on bilingual dictionaries and the main research effort was devoted to filter out the noise introduced by translation ambiguities. We experimented a disambiguation strategy based on Latent Semantic Analysis which allow us to compute the degree of semantic coherence of possible translation candidates. We also tested some query expansion methods in order to "balance" the fact that the system does not perform any kind of visual retrieval. Being the system basically the same, we are in a good position to compare results in the Image track with the one in the Ad-hoc track. Surprisingly it emerges that, contrary to what happened in the Ad-hoc exercise, expansion strategies improve results in image retrieval tasks.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; query formulation H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Italian, English*

General Terms

Languages, Measurement, Performance, Experimentation

Keywords

Italian, English, Query translation, Query expansion, Semantic disambiguation, Latent Semantic Analysis, Metasearch, Delegated search, Visual search,Image Search

1 Introduction

CELI (www.celi.it) is an Italian company active in the Natural Language Processing and Document Retrieval field. Over the years CELI developed several cross language information retrieval systems, both in the context of European projects and commercial applications. Recently an internal project started, with the goal of providing cross language access to the indexes of Google and Yahoo (cf. http://www.elois.biz). It is in the context of such a project that the participation to CLEF was decided. The goal was mainly to compute figures of performance of several different strategies of query disambiguation and query expansion as well as obtain a comparison with the best systems in the cross language information retrieval arena.

We were particularly interested to Image track for the following reasons:

- There is a renewed commercial interest for image search, and cross language image search is one of the few domains where appreciation of the results can take place irrespective of the query language.
- By working just on textual metadata the system is easily extensible to different kinds of digital objects, such as music and images.
- Thanks to a collaboration with the University of Bolzano, CELI is active in the domain of digital libraries, with a particular regard to the migration from standard OPAC search to natural language access to catalogues [1]. This justifies our interest for metadata based information retrieval.

2 System Description

The system which has been used is identical to the one described in [2] in this volume. Therefore we refer to such a paper for a detailed description. Here we just report the main features:

- Cross linguality is achieved via query translation, based on bilingual dictionaries.
- The typical user of the system is a standard web search engine user: therefore no domain restriction is imposed and no parametrization has been performed on the basis of the ImageCLEF corpus.
- In order to solve the traditional problem of multiplicity of translations we implemented a disambiguation strategy based on the semantic coherence of the resulting target query.
- Such a disambiguation strategy is based on semantic vectors associated to each translation. These have been computed by using Latent Semantic Analysis.
- Besides query translation, the system is able to perform query expansion on the target language. Three strategies are tested, namely:
  - The use of a *generic* semantic net.
  - The use of a “prioritized” semantic net that integrate domain specific extensions (legal, medical, geographical, etc.).
  - The use of “related concepts” as derived from the application of Latent Semantic Analysis.

3 Description of the Runs

Participation to ImageCLEF was aimed at testing the system in a completely new and unseen setting. CELI subscribed to the ImageCLEF experiment just few days before the deadline, so no tailoring of the system on the basis of the corpus was performed, nor any kind of experimental run using topics from previous years.

We submitted a total of 8 runs, each one with Italian topic descriptions (the `title` tag) targeting an English corpus of image metadata. From the IAPR corpus we indexed the `title`, `description`, `notes` and `location` fields. Out of eight runs, half of them put all translations in OR, whereas the remaining ones use AND among different translated query terms and OR to relate different translations and/or expansions of the same term. Each quartet of runs is then partitioned into the above mentioned query expansion strategies plus a a run with no expansion at all.

4 Results

As we stated in the introduction, the goal of the system is to act as cross language interface to mainstream search engines. It is not therefore surprising that, after research by [3], we consider as the main "index of success" precision as registered at the first 10 hits. In the following table we list the results of our run (ordered by precision at 10):

Run name	rel. retr.	Mean av. Pr.	Pr. at 10	Pr. at 20
CELI-AND_CwnExpansion	1474	0,1392	0,2233	0,2000
CELI-AND_NOEXPANSION	1551	0,1486	0,2150	0,2092
CELI-OR_NOEXPANSION	1551	0,1441	0,2033	0,1975
CELI-AND_CwnCascadeExpansion	1590	0,1362	0,1917	0,1817
CELI-OR_CwnExpansion	1474	0,1284	0,1917	0,1775
CELI-AND_LisaExpansion	1655	0,1237	0,1867	0,1725
CELI-OR_LisaExpansion	1655	0,1189	0,1633	0,1558
CELI-OR_CwnCascadeExpansion	1590	0,1208	0,1583	0,1492

The first important thing to notice is that AND based target queries usually outperform OR based ones. Therefore we exclude OR based runs from the analysis. Concerning the remaining ones it is interesting to see how they compare with respect to runs in the ad-hoc track, where, we repeat, the system in use was exactly the same:

Run name	rel. retr.	Mean av. Pr.	Pr. at 10	Pr. at 20
CELItitleNOEXPANSION	773	0,2397	0,2400	0,1890
CELItitleLisaExpansion	814	0,2238	0,2160	0,1720
CELItitleCwnCascadeExpansion	673	0,2390	0,2020	0,1650
CELItitleCwnExpansion	636	0,2110	0,1980	0,1520

Here we report results only for runs based on titles, as these share obvious similarities with ImageCLEF topics, for which, for Italian, only the title was available.

The first thing we notice is basically an average similar performance for precision at 10 and 20, in spite of a difference of about 10% in terms of mean average precision. Given the fact that rarely users inquiry beyond the 20th result, this is an indication that metadata-based cross language information retrieval can perform as well as full text information retrieval. This is in turn a good new for digital libraries for which only a small part of texts has been digitalized.

The second important thing is that while in the Ad-hoc track absence of query expansion seems to provide better results, in the case of image retrieval, the reverse applies. We think that this is an effect of the fact that metadata are usually shorter and the language is generally more controlled. It is therefore unlikely that synonyms might occur in the description of the same image. Query expansion will then intervene to bridge between the uncontrolled nature of the query and the controlled nature of the metadata.

Finally we observe that whereas LSA based expansion was ranked as the best expansion strategy in the Ad-hoc task, it performed very badly in the Image track. So far we do not have a clear explanation for this behavior. However we suspect that this is due to the fact that LSA expansion focuses on generic "related terms". These are likely to occur in documents centered around a given topic, but they are usually absent in the description of a picture. For instance if a news article is about a church, it might contain also the adjective *Christian* or the noun *religion*. By contrary, the description of a photo of a church is not likely to contain anything else apart from visually perceivable features.

5 Conclusion

The most evident limit of CELI's participation to ImageCLEF was of course the lack of any capability of performing visual retrieval. We estimate, however, that in the context of delegated search this is not a major drawback, as, in any case, mainstream image search engines are not likely, in the immediate future, to provide search restrictions based on visual patterns. If this were the case, we also estimate that standard web users would go on looking for content based features rather than visual features. A query for pictures of a "Mercedes E200 coupé" will probably be always more common than "a brown Jaguar on a green background".

In terms of retrieval based on metadata we estimate that big improvements are possible according to the following lines:

- Better recognition of named entities and evaluation if they should be translated or not.
- Expansion of named entities, in particular as far as geographical data are concerned.
- More selective use of metadata (currently they are just indexed as if their concatenation was the text of a document).
- Introduction of some basic reasoning capabilities. These might be considered as a sort of query expansion, but their implementation might vary. We are thinking to basic skeletons of temporal reasoning (an *old car* might have been indexed as a *1974 car*) as well as some ontology based inferences (A lion is an animal) .

Improvements on these points will be our goal for the next ImageCLEF experiment.

References

[1] Bernardi R., Calvanese D., Dini L., Di Tomaso V., Frasnelli E., Kugler U., Plank B. Multilingual Search in Libraries. The case-study of the Free University of Bozen-Bolzano *Proc. 5th International Conference on Language Resources and Evaluation - LREC 2006* (2006)

[2] Curtoni P., Dini L. CELI participation at CLEF 2006: Cross Language Delegated Search *CLEF 2006 Working Notes* (2006)

[3] Fallows D. Search Engine Users. *PEW INTERNET & AMERICAN LIFE PROJECT* (2005), www.pewinternet.org/pdfs/PIP_Searchengine_users.pdf