CELI participation at CLEF 2006: Cross Language Delegated Search

Paolo Curtoni
Luca Dini
CELI
Torino-Italy
curtoni/dini@celi.it

Abstract

In this paper we discuss the CELI's first year of activity at CLEF. 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 translation of queries with no direct access to the corpus, which basically 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 and we found that in general they do not increase the performance of the system. However,among the various adopted expansions, we found that the one based on LSA semantic grouping provided the best results. Our experiments are all about Italian topics targeting English documents.

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

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 lingual access to the indices 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 achieve figures on several different strategies of query disambiguation and query expansion as well as obtaining a comparison with the best systems in the cross language information retrieval arena.

In this paper we will describe the adopted methodology and we will try to compare the different approaches which have been followed. We will also stress the differences between the target application and the context of the experiment.

1.1 Cross Language Delegated Search

The goal of the application under development is to provide access to the web in the user language. Most specifically we want to enable any user to type a query in his/her own language and retrieve hits in any language. Reasons for performing cross language information retrieval on the web have been widely emphasized in [4], [9] and [7]. Here we just want to add that the evolution of the web has added new reasons why this could become more interesting than ever in the near future:

- E-commerce is providing excellent reasons for translating queries into different languages. A user might not be able to express the concept of "leather wallet" in Italian but be perfectly capable of making his/her choice in an Italian fashion web site. The same holds true for different kinds of equipments for which technical specifications are easy to read even in a completely unknown language.

- With the development of the multimedia web users might want to be able to access digital objects crosslingually. These are typically indexed with language-dependent metadata but, at the same time, they can be perfectly appreciated by users unaware of the metadata language. Images are the most obvious example, but also music (writing *Mozart concerto* is perfect for accessing music published by an Italian company, but not for a German or an English one) and, in certain cases, videos.

As the target document collection of the application is represented by the web the set up of an autonomous indexing machinery it is not an option, at least not for a company like CELI. Therefore the only viable approach is represented by delegated search. Simply stated, the user query is translated by our query translation module and issued to some mainstream search engine.[1].

This process (cross language delegated search) rises some issues that are sometimes negligible for standard CLIR.

- *Absence of a corpus*: no kind of parametrization is possible on the basis of the corpus, as the corpus is absolutely uncontrolled.

- *Lack of a domain*: the translation engine must be able to cope with words in any domain. Notice that this is different from having a "generic domain" such as the one found in newspapers: in our case both queries and documents might be very specific.[2]

- *Web spam*: In a controlled corpus usually all documents are there for the purpose of being retrieved. In the web there are many documents that are there with the purpose of cheating search engines.

- *kind of queries*: The queries which are issued to a generic search engine are different in style, syntax and accurateness from the ones issued to a controlled corpus. (c.f. [6] and [10].

Most of these issues will not be considered in this paper. However they are crucial to provide a rationale about certain features of our search engine, which might look peculiar in a context such as the CLEF experiment.

[1] On the web site www.elois.biz, in the case of Yahoo, result hits are provided by a web service and then processed by the calling application. In the case of Google, due to technical limitations of the available web service, results are served directly by Google (Adsense Program).

[2] The same absence of domain applies to cross language application in the domain of institutional digital libraries ([1])

2 System Description

2.1 DocDigger-CLIR

DocDigger is basically a multilingual search engine with plugins available for Italian, English, French, Polish and German. It is used in a number of commercial application (c.f. for instance http://www.comune.torino.it) and it is maintained by CELI since 5 years. In the CLEF experiment we have plugged the *CELI query translation module* into DocDigger in order to achieve cross-linguality. So basically the English target corpus is indexed monolingually and all cross language issues are dealt with in the query translation/expansion module.

2.2 Corpus Processing Steps

DocDigger has many processing functionalities including keyword/concept extraction, Named Entity extraction and indexing, date extraction, intelligent site conversion, etc. However, in the context of the CLEF experiment only the following processing steps have been considered as relevant:

- *Lemmatization* . Each word is translated into its morphological root and a category is assigned according to a part of speech disambiguation algorithm.[3]
- *Stop-words removal.* Stop-words are removed from the original text. These include all functional words as well as a manually compiled, domain independent, list of common words.
- *Indexing.* The indexing step was quite trivial: all the text in the in the documents was indexed. However document titles and document bodies have been indexed separately, in order to verify whether assigning an higher weight to query words contained in the document title would have produced better results.

2.3 Query Translation

The core of the system is represented by the query translation module. The approach is based on translation dictionaries available to CELI. Such dictionaries were created over the years as a resource stratification process. In particular they are made up of the following sources:

- Acquired commercial biligual dictionaries;
- Open source biligual dictionaries;
- Internal dictionary development;
- Bilingual domain specific thesauri;
- Translation computation by using machine learning methods applied on bilingual aligned corpora.

It goes without saying that such resources are quite rich but completely uncontrolled. In particular they suffer of the following problems

- Translations which are overtly wrong;
- Translations which are specific to certain domains (e.g. medical), thus extremely unlikely to be good translations in a generic context;
- Translations which are obsolete and no more (or scarcely) used in the target language.

[3] In this context the part of speech disambiguation algorithm is very poor. As syntactic disambiguation of input queries is usually not feasible, we just disambiguate according to the categories which are most likely to occur in queries.

pescare	catch, grab, take hold of, **fish**, catch, draw, capture, catch, get, gill
barca	small boat, **boat**, tub
mare	forest, timber, timberland, woodland, forest, wood, brine, blue,flood, tide, **ocean, sea**

Figure 1: Translation ambiguities for the Italian query *pescare barca mare*

This being the case, it is evident that for multiple words queries, such as the ones derived from CLEF topics, chances of providing a reasonable solution to the user request are extremely low. As an example, if we take for instance the Italian query *pescare barca mare*, which would be optimally translated as *fish boat sea* we will have to cope with the set of translations in figure 1.

In our application context (delegated search), if all the translations are retained, result hits may vary randomly according to the called search engine. For instance, for the above query the first hit of Google is www.worldseafishing.com, whereas for Yahoo! is www.timberland.com. Therefore the main goal of the research was, for multi terms queries, to get rid of contextually unreasonable translations by exploiting the semantic proximity of the terms in each possible combination of translations (in our example 297 possibilities).

2.3.1 Computing semantic proximity

In order to compute the semantic proximity of possible target translations we set up the following intermediate goals:

- Obtain a "semantic lexicon" which associates a vector of semantic features to each word.
- For each candidate target query[4] find a function which evaluate the degree of coherence of the query.
- Select the query (or the queries) which have the highest "coherency" score.

As an example,we can consider the Italian word *carta* which can be translated as *card, map, certificate, paper* or *document*. The goal is to select the *map* translation in the query *carta stradale*(en. *road map*), the *paper* translation in *carta stampante* (en. *paper printer*) and the *card* translation in *carta di credito* (en. *credit card*). This has to be achieved by comparing the semantic vectors associated to each possible translation sequence and selecting tuples of words with a minimal distance (see Figure 2).

As no sufficiently rich manually encoded thesaurus was available, we decided to induce semantic vectors for words from corpora. As the objective of the project is to provide a cross language interface to the web, the corpora should have been neither standard balanced ones nor domain specific ones: we needed corpora reflecting the information available on the web in a balanced way. Fortunately we could obtain balanced web corpora in different languages from the WaCky project ([2], [8]). Due to processing times, we randomly sampled from such corpora a smaller learning corpus of about 700 M of pure text per language: we estimate this to be a sufficiently large snapshot of what the web is offering. The corpora were then lemmatized, POS tagged and cleaned of stop-words, according to the same criteria adopted for document and query processing.

As for the learning of semantic vectors we adopted a technique which is very close to Latent Semantic Analysis ([5]). We decided to learn semantic vectors for about 30.000 words. 1000 words where automatically selected for the initial matrix population and a textual window of 7 words was considered. Each word has been assigned about 100 semantic features[5], which were obtained after 500 iteration of the singular value decomposition algorithm.

With the semantic lexicon in place, the algorithm for selecting the best translation is quite easy: we just compute the Cartesian product of the translation of each source term ad we obtain

[4] In the context of this paper the expression "candidate target query" is used to mean one of the combinations deriving from the translation of source terms. For instance if the source query is $< ws_1\, ws_2 >$ and each word has two translations, the set of candidate target queries is $\{< wt_1^a\, wt_2^a >< wt_1^b\, wt_2^b >< wt_1^a\, wt_2^b >< wt_1^b\, wt_2^a >\}$.

[5] In the context of LSA features are just doubles associated to a position in a vector. For each word the value of a certain position in the associated vector represents the degree "proximity" of that word to that feature.

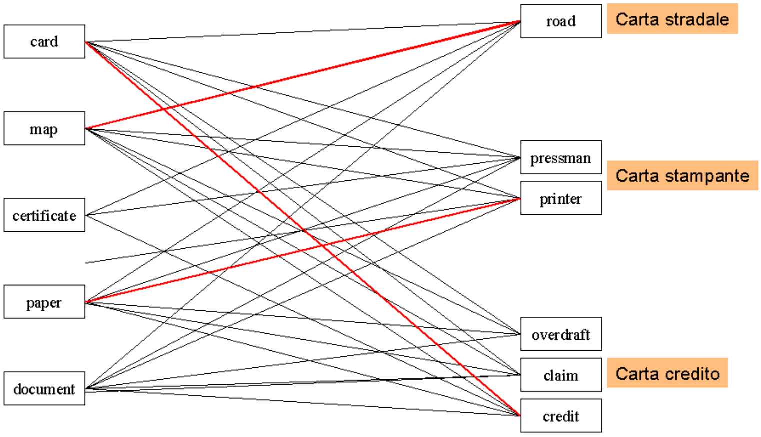

Figure 2: Possible combinations of translations of *carta* associated with possible translations of *strada, stampante and credito*. Red lines indicates semantic proximity.

a set of target candidate queries. For each candidate we then apply a scoring function which, on the basis of the semantic vector associated to each translation, computes the proximity of words in the translation. Such a function is assumed to tell us that, for instance, in our example 1 the candidate *gill boat forest* is "less cohesive" than *fish boat sea*, thus a worse translation of the initial query. The function for computing the cohesion for two terms is the following one (dot product):

$$Neighborhood(V1, V2) = \sum_{0 \leq i < 100} weight(V1, i) * weight(V2, i) \qquad (1)$$

If the candidate query is composed of more tha two words, we just apply the neighborhood function pairwise to each couple of possible candidates.[6]. Eventually we select the candidate for which the neighborhood function return a maximal value.

2.4 Query Expansion

Once the best candidate is obtained, in the context of the CLEF experiment, we have experimented several strategies of query expansions:

Wordnet expansion: a semantic net similar to Princeton Wordnet is applied in order to expand *target* query terms.

Wordnet cascade expansion: several semantic nets are used, with different orders of priority. The basic idea, here, is to introduce expansions which are not only derived from standard language, but also from technical jargons (medical, legal, mechanical, etc.)

LISA expansion: in this case expansion is driven by the semantic vectors which have been described in the previous section. Basically, for each term we expand it into the terms which are closer in terms of the neighborhood function (c.f. also [11]). For CLEF experiments we decided to use at most 5 expansions for each term, provided they had a proximity value higher than 0.3.

Expanded terms are always added as OR terms to the original term.

[6] It should be noticed that the selection of the best candidate is actually a bit more complex. Indeed we have to face the problem which occur where two equally unlikely terms in a candidate query show a high degree of cohesion. For instance if we have among the unlikely translation candidates of *mare* the word *wood* and if the original query is *veliero tre alberi mare* (en. *sailing boat three mast sea*)we want to avoid that the (contextually wrong) translation *tree* of *albero* "attract" *wood* as a translation for *mare*. By using, again, a latent sematic analysis technique, we dispose for each pair of words w1 ad w2 in two languages L1 and L2 an index of likelihood that the word w1 is a translation of w2. Such an index is taken into account in selection the best translation candidates

3 Description of the Runs

Participation to CLEF was aimed at testing the system in a completely new and unseen setting. CELI subscribed to the CLEF 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.

Due to resource constraints only the ad hoc experiment was performed, with Italian topics translated into queries in order to target English documents. Overall 12 runs have been submitted, with the following criteria:

- six runs use queries derived from titles, the remaining six use queries derived from descriptions.
- out of each group of six runs, three of them use the expansion methods described in the previous section. A run without expansion was always submitted
- The remaining two runs use a retrieval algorithm which gives more importance on document titles than document text. This boost factor was applied for a run with no expansion and for a run where terms were expanded using latent semantic analysis.

4 Results

As we stated in the introduction, the goal of the system is to act as cross-lingual 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 compare the results of our run (ordered by precision at 10):

Run name	rel. retr.	Mean av. Pr.	R Pr.	Pr. at 10	Pr. at 20
CELItitleNOEXPANSION	773	0,2397	0,2381	0,2400	0,1890
CELIdescNOEXPANSION	764	0,2268	0,2381	0,2320	0,1720
CELItitleLisaExpansion	814	0,2238	0,2212	0,2160	0,1720
CELItitleCwnCascadeExpansion	673	0,2390	0,2400	0,2020	0,1650
CELItitleCwnExpansion	636	0,2110	0,2074	0,1980	0,1520
CELItitleNOEXPANSIONboost	680	0,2035	0,2036	0,1900	0,1430
CELIdescLisaExpansion	793	0,1941	0,2016	0,1840	0,1470
CELIdescCwnExpansion	594	0,1792	0,1900	0,1760	0,1460
CELIdescCwnCascadeExpansion	602	0,1957	0,1982	0,1720	0,1380
CELIdescLisaExpansionboost	767	0,1908	0,1966	0,1660	0,1370
CELIdescNOEXPANSIONboost	733	0,1812	0,1811	0,1600	0,1300
CELItitleLisaExpansionboost	712	0,1732	0,1795	0,1600	0,1370

It can be noticed that best results on the first 10 hits are obtained by performing no expansion at all. If expansion has to be performed, than the best results are obtained by using the expansion which exploits semantic groping computed by means of LSA rather than manually coded semantic resources. On this respect it could also be noticed that best expansion results might have been obtained if semantic vectors had been computed on the basis of a corpus of news, rather than a web one.

Boosting on document title does not usually provide any positive effect. We also notice that the systems performs better on queries obtained from topic titles than from topic descriptions. This is a consequence of the fact that it is designed to handle web queries, i.e. queries usually containing few terms.

5 Conclusion

There are many points in which the system could be improved. Here we do not consider, of course, improvements in terms of document indexing and retrieval strategies, as the objective of the research is rather the one of exploiting the retrieval capabilities of mainstream search engines. However, in terms of query translation we noticed that many of the least precise queries (those with an R-Precision value lower than 10%) were affected by the following problems:

- Improper recognition of proper names and failure to understand whether they should have been translated or not.
- Lack of one or more translations in the dictionary.
- Presence among translation possibilities of rare and uncommon words which, under a standard TF*IDF retrieval strategy tend to have negative influences on document ranking.
- failure to recognize compound words in the target language (e.g. *specie in via di estinzione* → *engendered species*)

Besides improving on these aspects, the system would also benefit of a more accurate tuning of the expansion capabilities. In particular query expansion should be activated selectively and possible on the basis of some "a priori" index of the precision of the target query. This will be our goal for the next CLEF experiment. However we estimate that, as it stands, the system is already able to provide a reasonable help to user willing to perform cross language search on the web. In the last months of 2006 the www.elois.biz web site will then be appropriately advertised.

References

[1] Bernardi R., D. Calvanese, L. Dini, V. Di Tomaso, E. Frasnelli, U. Kugler, B. Plank 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] Ciaramita M., Baroni M. A Figure of Merit for the Evaluation of Web-Corpus Randomness *Proceedings of EACL 2006* (2006), 217–224.

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

[4] Gey F.C., N. Kando, C. Peters Cross language information Retrieval: a research roadmap. *SIGIR Forum, 36 (2)* (2002), 72–80.

[5] Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: The Latent Semanctic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review , 104* (1997), 211–240.

[6] Jansen, B.J., Spink, A., Saracevic, T. Real life, real users and real needs: A study and analysis of users queries on the Web. *Information Processing & Management, 36(2)* (2000), 207–227.

[7] Petrelli D., S. Levin, M. Beaulieu, M. Sanderson Which user interaction for cross-language information retrieval? Design issues and reflections *Journal of the American Society for Information Science and Technology archive, 57 (5)* (2006), 709–722.

[8] Sharoff, S. ating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus* (2005),

[9] Sigurbjornsson B., J. Kamps, M. de Rijke Blueprint of a Cross-Lingual Web Retrieval Collection *JOURNAL OF DIGITAL INFORMATION MANAGEMENT 3 (4)* (2005),

[10] Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T. Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology, 52(3),* (2001), 226–234.

[11] D. Stenmark, Query Expansion on a Corporate Intranet: Using LSI to Increase Precision in Explorative Search *Proceedings of the 38th Hawaii International Conference on System Sciences - 2005* (2005).