Searching Cross-Language Metadata with Automatically Structured Queries

Víctor Peinado, Fernando López-Ostenero, Julio Gonzalo, and Felisa Verdejo

NLP Group - UNED. ETSI. de Informática,
28040 Ciudad Universitaria, Madrid, Spain
{victor, flopez, julio, felisa}@lsi.uned.es
http://nlp.uned.es

1 Introduction

When searching metadata, it can be useful to detect expressions in the query that should be searched for in specific fields (for instance, person names might correspond to an "author" field). In [1], it was shown that automatically structured queries (matching title, abstract, author and publication fields) improved effectiveness when searching the ACM, CITIDEL and NDLTD Computing Digital Libraries.

In a cross-language retrieval setting, we can decide, in addition, how to translate different types of information (named entities, temporal references, quantities, etc.) once they are automatically detected in the query.

This is the approach to cross-language metadata search that we test in this paper. We experiment with a simple strategy that **i)** locates proper names, temporal references and numbers in the query; **ii)** attempts to classify them by checking whether they appear as "author", "location" or "date" in the collection; **iii)** uses positive cases to structure the query, forcing the search engine to favour documents with the appropriate author, location or date.

2 Experimental Settings

We have used the Spanish-English version of the ImageCLEF ad hoc task testbed. In order to locate and identify named entities, temporal references and numbers appearing in the ImageCLEF queries, we have used a simple strategy that was enough for our purposes (see [2] for further details).

Then, we have compared three approaches: **i**) a **naive baseline** using a word by word translation. **ii**) a **strong baseline** following Pirkola's proposal [3], where alternative translations for a query term are taken as synonyms, giving them equal weights, and; **iii**) our **structured query approach**, which incorporates **field search** operators in addition to Pirkola's strategy. All three conditions have been tried with six different bilingual dictionaries.

Besides, we have evaluated three additional runs for comparison purposes: two monolingual runs (a straight run with the English version of the query, and an enhanced run with the field search strategy described in [2], and an additional

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 529–530, 2005.

cross-language run where named entities and temporal references are annotated manually. The latter is intended to evaluate the effects of errors in the automatic location of entities.

3 Results and Discussion

The results of the experiment are shown in Table 1. For all bilingual dictionaries, our structured query approach is better than the naive and Pirkola baselines. Pirkola's approach is, in turn, substantially better than the naive run in all cases. Only the differences between our structured query approach and the naive baselines are relevant according to a non-parametric Wilcoxon sign test (in half of the cases). Our best runs achieve an average precision of .54, which represents 91% of the best monolingual run (*monolingual+field search*). This result slightly outperforms the best official cross-language run in the ImageCLEF 2004 evaluation (which was .53, obtained by Dublin City University with the DE→EN language pair).

Remarkably, the manual annotation of named entities and temporal expressions does not improve the results obtained with our simple automatic recognition strategy. This is an indication that the field search strategy is reasonably robust: for instance, if an expression is misinterpreted as a person name, it will probably not appear in the author field and therefore precision will hardly be affected.

Table 1. Experimental results. Non-interpolated mean average precision (MAP) for different combinations of retrieval strategy and bilingual dictionary. "*" denotes a statistically significant difference with respect to its naive counterpart.

Dictionary	naive	Pirkola	field search	Additional reference runs	MAP
FreeDict	.34	.38	.42		
EWN	.36	.50	.52*	Monolingual base	.57
EWN2	.38	.51	**.54***	Monolingual+field search	.59
Vox	.40	.45	.53	CL manual field search	.54
All-Vox	.34	.52	**.54***	Best CL ImageCLEF run	.53
All	.37	.49	.53		

References

1. M. A. Gonçalves and E. A. Fox and A. Krowne and P. Calado and A. H. F. Laender and A. S. d. Silva and B. Ribeiro-Neto. The Effectiveness of Automatically Structured Queries in Digital Libraries. JCDL 2004
2. V. Peinado and J. Artiles and F. López-Ostenero and J. Gonzalo and F. Verdejo. UNED at Image CLEF 2004: Detecting Named Entities and Noun Phrases for Automatic Query Expansion and Structuring. Cross Language Evaluation Forum, Working Notes for the CLEF 2004 Workshop (2004)
3. A. Pirkola. The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. SIGIR'98 (1998) 55–63