Searching Cross-Language Metadata with Automatically Structured Queries

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

ETSI de Informática - UNED
c/ Juan del Rosal, 16
E-28040 Madrid, Spain

NLP Group

{victor, flopez, julio, felisa}@lsi.uned.es
http://nlp.uned.es

1 Introduction

- When searching metadata, it's useful to detect expressions in the query that should be searched for in specific fields (person names might correspond to an "author" field).
- [2] showed that automatically structured queries improved effectiveness when searching Digital Libraries.
- In a cross-language retrieval setting, we can decide how to translate named entities (proper nouns, temporal references, quantities) once they are automatically detected in the query.

2 Experimental Settings

ImageCLEF 2004 ad-hoc Testbed [1]

- 25 topics written in Spanish.
- 28,133 photographs annotated with rich semi-structured captions. Image descriptions consist of eight metadata fields (a unique ID, short and long titles, location, description, date, author, classification).
- a pool of relevance judgements generated at the track.

Named entity recognition

Proper nouns: expressions in uppercase wherever uppercase is not prescribed by punctuation rules.

Temporal references: expressions matching words such as weekdays, months or seasons.

Numbers: any numerical expression or words from a given list

For each entity located in the Spanish topic titles:

1. If it's a named entity, we ask the search engine to find any document containing the entity in the "author" or "location" fields. If the search is non-nil, we assume that the role of the entity is the field in which it was found.
2. If it's a cardinal number, we ask the search engine to find any document containing the entity in the "date" field. If the search is non-nil, we assume that the cardinal number represents a date.
3. If it's a temporal reference, we check if it's a date with a similar procedure.

Three approaches

naive baseline: using a word by word translation. Words which are not present in the bilingual dictionary are left untranslated.

strong baseline: following Pirkola's proposal [3], where alternative translations for a query term are taken as synonyms, giving them equal weights.

field search: our structured query approach, which incorporates field search operators in addition to Pirkola's strategy.

3 Experiments

We tried all three approaches with six different bilingual dictionaries:

FreeDict: a freely available online dictionary.

EWN: generated from the official EuroWordNet multilingual semantic network [4].

EWN2: compiled from an updated version of the Spanish Wordnet.

Vox: an electronic version of the Vox-Harraps ES→EN dictionary.

All-Vox: a combination of all the dictionaries above except Vox.

All: a merged version of all four dictionaries.

We evaluated three additional runs for comparison purposes: two monolingual runs (a straight run with the English version of the query, and an enhanced one with our field search strategy) and one additional cross-language run where named entities were annotated manually, in order to evaluate the effects of errors in the automatic location of entities.

4 Results and discussion

- For all bilingual dictionaries, our structured query approach is better than the naive and Pirkola baselines.
- Pirkola's approach is 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 (91% of the best monolingual run (*monolingual+field search*).

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

5 Conclusions

- Automatic query structuring seems an effective strategy to improve cross-language retrieval on semi-structured texts.
- No sophisticated named entity recognition machinery is required to benefit from query structuring.
- It remains to be checked whether this result holds on collections with different metadata fields and different textual properties.

[1] P. Clough, M. Sanderson and H. Müller. The CLEF Cross Language Retrieval Task (ImageCLEF) 2004. *Results of the CLEF 2004 Evaluation Campaign*. LNCS 3491. Springer Verlag. 2005.

[2] M.A. Gonçalves, E.A. Fox, A. Krowne, P. Calado, A.H.F. Laender, A.S.d. Silva, B. Ribeiro-Neto. The Effectiveness of Automatically Structured Queries in Digital Libraries. *Joint Conference on Digital Libraries (JCDL 2004)*. 2004.

[3] A. Pirkola. The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. *SIGIR'98*. pp. 55–63. 1998.

[4] P. Vossen. Introduction to EuroWordNet. *Computers and the Humanities, Special Issue on EuroWordNet*. 1998

Acknowledgements: This work has been partially supported by the Spanish Government under project R2D2-Syembra (TIC2003-07158-C04-02) and a PhD grant by UNED.