UNED at Image CLEF 2004: Detecting Named Entities and Noun Phrases for Automatic Query Expansion and Structuring

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

Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia (UNED)
c/ Juan del Rosal, 16, Ciudad Universitaria, 28040 Madrid - Spain
{victor}@lsi.uned.es

Abstract. This paper describes UNED experiments at the Image CLEF bilingual ad hoc task. Two different strategies are attempted: i) automatic expansion and translation using noun phrases; ii) automatic detection of named entities in the query for structured search on image caption fields.

All our experiments obtain results above the average MAP for the bilingual task. Structured searches using named entities improve performance over a strong baseline (Pirkola's structured query approach), achieving one of the best results for the whole bilingual track. Expansion with noun phrases, however, degrades results, possibly due to the mismatch between train and test collections.

1 Introduction

For its first participation in the Image CLEF task, the UNED NLP & IR Group took part in the bilingual ad hoc retrieval task, using Spanish and English as source and target languages, respectively. As in the the classic TREC ad hoc task, the main goal was, given a set of topics in a source language, to retrieve as many relevant images as possible from a collection in the target language.

Participants were provided with a list of topic statements and a collection of images with semi-structured captions in English. Every topic consisted of a title (a short description of the required search in few words) and a narrative (a description of what constituted a relevant and a non-relevant image for the search). Narrative was not provided for Spanish, hence our experiments use only the title field. The collection comprises 28,133 photographs from one of the most important sets of historic photography in Scotland.[1] All images have an accompanying textual description consisting of 8 distinct fields (e.g. a unique ID, both short and long titles, the location, a description of the image, the date, the author and some categories in which the photograph may be included). This rich meta information was the basis to retrieve relevant images in our approach.

We experimented with two different strategies:

[1] See http://ir.shef.ac.uk/imageclef2004/stand.html for further details.

1. Expand queries with noun phrases. Queries are expanded with related noun phrases, looking up a bilingual Spanish-English noun phrase list which was extracted from the CLEF news comparable corpus (LA Times 1994, Agencia EFE 1994).
2. Identify named entities and dates in queries, and perform structured queries against appropriate image caption fields:
   a) Proper names are searched in the "author" and/or "location" fields. If the search is non-nil, the retrieval mechanism favours images containing these entities in that fields.
   b) Temporal references are searched in the "date" field. If the search is non-nil, images matching the temporal reference in the date field are favoured.

Prior to experimenting with these strategies, we first built an improved English-Spanish translation resource, merging our in-site dictionaries with free web resources.

Reasons to try the above techniques include:

1. Expansion with Natural Language approaches is more benefitial with short queries [1] or short documents such as image captions [2]. Being in a Cross-Language search context, we decided to experiment with our query expansion technique based on aligned noun phrases, that gave excellent results in the CLEF interactive track [3–5].
2. Since the ImageCLEF collection contains structured image captions (including author, date, location and description fields) it seems interesting to explore the possibility of detecting different types of information in the query to perform more precise searches. Then, we experiment with a simple strategy that tries to match every named entity as a possible author name or location.

This paper is structured as follows: in Section 2, we first discuss the possibilities of using noun phrases in query expansion and interactive tasks, then we present the linguistic resources used (Section 2.1), our preliminary CLIR experiments on the CLEF collection (Section 2.2) and the settings of our Image CLEF experiments (Section 2.3). Then, we explain the structured searches approach (Section 3) and discuss the official results obtained in the track (section 4). Finally, in Section 5, we draw some conclusions.

2 Query expansion with noun phrases

2.1 Linguistic resources

We used two comparable corpora from the CLEF ad-hoc track: the Spanish newswire collection EFE 1994, and the Los Angeles Times 1994 news collection. Out of these comparable corpora, we built a bilingual dictionary containing more than five million aligned noun phrases[2]. Noun phrases are automatically recognized and extracted using statistical data such as the frequency of sequences of two or three informative words (nouns and adjectives) in both languages [6]. We consider that two phrases are aligned if they have the same amount of informative words and there is a one-to-one correspondence using a bilingual dictionary (see Table 1 for examples).

English	**Spanish**
Orange County	Condado de Orange
abortion issue	tema del aborto
free trade agreement	acuerdo de libre comercio
World War II	Segunda Guerra Mundial

Table 1. Example aligned phrases using CLEF comparable corpora.

Previous UNED participations in the iCLEF track[3] proved the utility of noun phrases for document selection [3], query translation and refinement ([4, 5]).

Finally, we built a new bilingual Spanish-English dictionary made up from heterogeneous lexicographic resources such as dictionaries and word lists (some of them freely available in the Web) and semantic networks such as WordNet [7] and EuroWordNet [8]. Every source went through a cleaning process before merging them in an XML-structured dictionary showing up all the information from each original source, almost without typos or inconsistencies. As a result of this merging, our final dictionary contains more than 57,000 entries in Spanish and 85,000 in English, with a total size of about 50 Mb[4].

2.2 Preliminary experiments over the CLEF collection

In order to check the usefulness of noun phrases for Cross-Language ad hoc retrieval, we have performed a number of experiments with 140 Spanish CLEF topics (corresponding to 2001-2003 campaigns) and the LA Times 1994 English CLEF collection. We start with the three following baselines based on word by word translation using bilingual dictionaries:

naive baseline Word by word translation, building a bag of words with all the possible translations appearing in our dictionaries.

[2] Roughly, there are more than 4.5 million phrases containing two informative words and 850,000 containing three.

[3] The Interactive track for the Cross-Language Evaluation Forum webpage is available at `http://nlp.uned.es/iCLEF`.

[4] Using this merged dictionary instead of the original VOX dictionary we used in previous approaches, there is an improvement in CLIR experiments with the CLEF collection of 36%

frequencies We built a bag of words from only those possible translations appearing in more than one lexicographic source, assuming that they should be the most common and reliable.

Following this strategy, we pursued two goals: on one hand, we used only those translations considered reliable. On the other hand, we rejected residuary translations from semantic networks in order to reduce the noise produced by the expansion.

strong baseline We used Pirkola's proposal [9] to build structured translated queries, using the synonymy operators implemented in the INQUERY search engine [10] to wrap alternative translations for every word in the query.

systran The queries were translated using the Systran machine translation system.

These baselines are compared with three runs using the bilingual noun phrase list:

phrases + pirkola We used our noun phrases dictionary to expand the query with related noun phrases and translate them using a bilingual dictionary. Our strategy was the following: firstly, we expanded each topic term with the ten most frequent noun phrases containing the term in the CLEF collection and then we translated the phrases using the aligned noun phrase list. Those query terms from which no phrases were identified were included in the translated query using Pirkola's approximation, i.e. using synonymy operators.

"multi-lemma" phrases + pirkola In order to limit the noise produced by the phrase expansion of the previous experiment, we only use noun phrases containing at least two query terms.

phrases + pirkola + systran Combined run using all three resources, i.e. noun phrases, structured translations using our dictionaires and the Systran machine translation system.

As shown in Table 2, our experimental proposals using noun phrases outperformed the baselines. The differences were statistically significant according to a non-parametric Wilcoxon sign test. The strong baseline obtained the same average precision than Systran's translations, showing that the combination of Pirkola's structured query approach with reliable lexicographic resources is an excellent CLIR baseline.

In summary, our results show that:

– Phrases do improve CLIR results, at least when the training corpus and the test corpus are similar. Even though the porcentual gain is not very high, in a setting with very small documents (e.g. image captions or topic titles) it would be reasonable to expect higher improvements.
– There is no need to use external machine translation systems, at least when translating small documents.
– The "multi-lemma" variant of noun phrase expansion performs slightly better for batch CLIR, although the difference is not statistically relevant according to a Wilcoxon sign test.

run	Avg. precision
naive baseline	.19
frequencies	.25
strong baseline	.27
systran	.27
phrases + pirkola	.29
phrases + pirkola + systran	.30
‘‘multi-lemma’’ phrases + pirkola	**.31**

Table 2. Results of our preliminary experiments

– The quality of a translation strongly depends on the resources used.

2.3 Settings for ImageCLEF experiments

Extending the above results to the Image CLEF bilingual ad hoc task, we have used the corpus and the set of Spanish topics provided by the organization, our bilingual XML dictionary, the Systran machine translation system[5] and the set of aligned noun phrases between English and Spanish.

Topic titles were processed, stopwords and punctuation removed[6] and content words lemmatized before translation using the merged dictionary or the noun phrase bilingual list.

Since the image captions contain structured information, we decided to use it by identifying which query terms could be understood as authors, locations or dates.

3 Structured search using image caption fields

3.1 Entities Recognizer

We used a set of simple rules to identify named entities, temporal references and numbers in the queries:

Named entities Expressions in uppercase wherever uppercase is not prescribed by punctuation rules.

Temporal references Those ones matching words such as names of weekdays, months or seasons.

Numbers Those ones matching any numerical expression or words from a given list (e.g. *dos* (2), *cien* (100), *mil* (1,000) . . .)

[5] Systran web-based interface available at `http://www.systransoft.com`

[6] In order to adapt the stopword list to this specific task, we included as stopwords *fotografías*, *fotos* (photographs), *retrato* (portrait). . .

3.2 Structured search over image caption fields

For each entity located in the Spanish topic titles:

- If it is a named entity, we ask the search engine to find any document containing the entity in the "author" or "location" fields, first in Spanish and then in English.[7] If the search is non-nil, we assume that the role of the entity is the field in which it was found.
- If it is a cardinal number, we ask the search engine to find any document containing the entity in the "date" field. If the search in non-nil, we assume that the cardinal number represents a date.
- If it is a temporal reference, we check if it is a date, in the same fashion.

3.3 Entities, dates and numbers found in the queries

In Table 3, we show the entities found for each Spanish topic title. Our recognizer located 31 entities (named entities, temporal references and cardinal numbers), although some of them are incorrect. For instance, on topic 5 *Irlanda* and *Norte* should have been identified as a unique named entity *Irlanda del Norte*. *Elisabeth*, on topic 14, was not identified as a possible entity. Besides, on topics 11 and 13, expressions such as *Postales* and *Campeonato Abierto* were misidentified as named entities.

Entities such as *Postales*, *Campeonato Abierto*, *Reina Madre* and *Segunda Guerra Mundial* did not represent any author, location or date. In any case, our strategy did not identify them as such either.

Regarding the other located entities, a manual analysis about their roles showed that:

authors Every possible author (*Thomas Rodger*, *John Fairweather* and *George Middlemass Cowie*) was correctly identified using this strategy.

locations *Roma*, *Irlanda*, *Norte*, *British Columbia*, *Canadá*, *Egipto*, *Londres*, *Bute*, *Escocia* and *York* were correctly identified. *St. Andrews*, *Cambridge*, *Tay Bridge*, *Crail Camp*, *North Street* and *Edimburgo* were not.

dates *Abril*, *1908*, *1879*, *1939*, *1954*, *1900.* All dates were identified with this strategy.

Overall, the algorithm is reasonably precise, given the very simple heuristic rules used for detection. But there is still room for improvement using proper Named Entity Recognizers.

[7] We perform the search in both languages because there is no general rule for translating proper names. Entities were translated using Systran because of the lack of proper names in our dictionary.

topic #	Entities
1	Retratos de ministros de la iglesia por [$_{NE}$ Thomas Rodger].
2	Fotos de [$_{NE}$ Roma] que fueron tomadas en [$_{DATE}$ Abril] de [$_{CARD}$ 1908].
3	Vistas de la catedral de [$_{NE}$ St. Andrews] por [$_{NE}$ John Fairweather].
4	Hombres vestidos militarmente, [$_{NE}$ George Middlemass Cowie].
5	Buques de pesca en [$_{NE}$ Irlanda] del [$_{NE}$ Norte].
6	Vistas panorámicas en [$_{NE}$ British Columbia], [$_{NE}$ Canadá].
7	Vistas exteriores de templos en [$_{NE}$ Egipto].
8	Edificios de la universidad o colegios universitarios, [$_{NE}$ Cambridge].
9	Fotos de faros ingleses.
10	Calles en plena actividad en [$_{NE}$ Londres].
11	Tarjetas [$_{NE}$ Postales] con múltiples vistas de [$_{NE}$ Bute], [$_{NE}$ Escocia].
12	Desastre ferroviario en el [$_{NE}$ Tay Bridge], [$_{CARD}$ 1879].
13	Torneo del [$_{NE}$ Campeonato Abierto] de golf, [$_{NE}$ St. Andrews] [$_{CARD}$ 1939].
14	Elizabeth la [$_{NE}$ Reina Madre], en su visita a [$_{NE}$ Crail Camp], [$_{CARD}$ 1954].
15	Daños provocados por bombas en la [$_{NE}$ Segunda Guerra Mundial].
16	Fotos de la catedral del [$_{NE}$ York].
17	Vistas de [$_{NE}$ North Street], [$_{NE}$ St. Andrews].
18	Fotos del castillo de [$_{NE}$ Edimburgo] antes de [$_{CARD}$ 1900]
19	Gente marchando o desfilando.
20	Río con un viaducto al fondo.
21	Monumentos a los caídos en la guerra en forma de cruz.
22	Fotos mostrando tradicionales bailarines escoceses.
23	Fotos de cisnes en un lago.
24	Golfistas golpeando con sus palos de golf.
25	Barcos en un canal.

Table 3. Named entities, temporal references and cardinal numbers located for each topic title.

4 Results and discussion

4.1 Submitted runs

Given the preliminary results discussed in Section 2.2, we decided to use the following strategies in our ImageCLEF experiments:

- Naive baseline using a word by word translation (`UNEDESBASE`). For instance, topic 13 (*Torneo del Campeonato Abierto de Golf, St. Andrews 1939)* produces:
  ```
  topic 13: turn tourney tournament tourney joust tilt championship title
  open frank open-minded opened overt unconcealed undone up extrovertish
  unfastened unlatched unlocked unsecured exposed hospitable forthright
  open-ended unresolved outgoing assailable undefendable undefended unhealed
  open undo dig head lead blossom unlock spread unfold brighten clear golf
  st andrews 1939
  ```
- Strong baseline using a structured query, following Pirkola's approach (`UNEDES`).

This is the core of the structured query for the next approaches, using INQUERY's synonymy operators:

```
topic 13: #syn( turn tourney ) #syn( tournament tourney joust tilt )
#syn( championship title ) #syn( open frank open-minded opened overt
unconcealed undone up extrovertish unfastened unlatched unlocked
unsecured exposed hospitable forthright open-ended unresolved outgoing
assailable undefendable undefended unhealed ) #syn( open undo dig head
lead blossom unlock spread unfold brighten clear ) golf st andrews 1939
```

- Structured query using INQUERY's operators and structured search over captions (`UNEDESENT`).
  If some entity is located and identified as a possible author name, location or date, we include the structured search over the caption fields. In this case, the search engine will favor those images in whose caption fields *1939* is tagged as a date. So, the following operator is added to the previous query:
  `#field( DDATE #sum( 1939 ) )`
- Structured query using INQUERY's operators and structured search over captions + noun phrases (`UNEDESENTNOO` and `UNEDORENTNOO`).
  We detected several errors in the original Spanish query set. These were fixed and sent to ImageCLEF organizers for distribution among other participants. However, for completeness, we submitted the most complex runs both with the original and the fixed query set (`UNEDORENTNOO` and `UNEDESENTNOO`, respectively).
  In order to expand the queries, we added the set of noun phrases extrated from the query terms using the "multi-lemma" phrases strategy. For topic 13 and `UNEDESENTNOO`, the phrases included are:

  ```
  #phrase( golf course manager ) #phrase( world golf championship )
  #phrase( world championship tournament ) #phrase( first golf tournament
  )
  #phrase( day after a golf tournament ) #phrase( chiefs into the title )
  #phrase( clear the tournament ) #phrase( champions tournament )
  #phrase( conference tournament title ) #phrase( day golf tournament )
  #phrase( golf tournament last ) #phrase( league golf tournament )
  #phrase( tournament of champions ) #phrase( phoenix golf tournament )
  #phrase( tennis tournament in st ) #phrase( championship golf course )
  #phrase( ups for golf ) #phrase( championships golf tournament )
  #phrase( gains after a bond ) #phrase( final of the tournament
  of champions ) #phrase( tournament at st ) #phrase( bond gains )
  #phrase( title of chief )
  ```

Summing up, the set of submitted runs and its features are shown in Table 4.

4.2 Results

All five runs obtained results above the average (average MAP for Spanish as query language is 0.30). Our best run, `UNEDESENT`, was the best Spanish →

run	word translation	noun phrases	structured caption
UNEDESBASE	bag of words	X	X
UNEDES	Pirkola	X	X
UNEDESENT	Pirkola	X	√
UNEDESENTNOO	Pirkola	√	√
UNEDORENTNOO	Pirkola	√	√

Table 4. UNED submitted runs.

English submission, 88% of the best monolingual run and 97% of the best cross-language submission (DCU German → English). Our results are shown in Table 5.

Structured queries over image captions (UNEDESENT) obtained an improvement of around 8.3 % with respect to Pirkola's approach (UNEDES). Apparently, a very simple detection of entities can be useful to improve retrieval results using the rich structure of image metadata.

Expansion with noun phrases does not improve over the baseline. The main reason is that our set of aligned noun phrases had been previously extracted from a collection of very different genre (newswire). As shown on section 4.1, the expansion inserted too much noise to get better average results.

Finally, it is worth noticing that our weakest baseline (UNEDESBASE) is about 26% better than the average MAP for the Spanish participants, confirming that a good bilingual resource is at least as important as the CLIR technique being used.

run	MAP	% monolingual
Best monolingual	0.59	-
Best cross-language	0.53	90
Best ES - EN	0.52	88
UNEDESENT	0.52	88
UNEDES	0.48	82
UNEDESENTNOO	0.47	80
UNEDORENTNOO	0.42	72
UNEDESBASE	0.38	64
average	0.30	50.87

Table 5. Results for UNED runs.

5 Conclusions

In this paper, we have presented two different strategies applied to the ImageCLEF bilingual ad hoc task:

– Expand queries with noun phrases, translating and expanding the queries with noun phrases automatically extracted from a different corpus. This expansion degrades retrieval results in our experiments, indicating that techniques based on bilingual comparable corpora can be useless when the training and test domains are very different.
– Perform structured searches using named entities and dates automatically located in the query. This technique obtains an improvement of 8.3 % with respect to our baseline (Pirkola's approach) and can be easily extended to other searches over structured documents.

In addition, even our simplest baselines have performed above the average, showing that work on merging bilingual dictionaries can be as important as the retrieval strategy in CLIR tasks.

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project R2D2 (TIC2003-07158-C0401), and UNED Ph.D. grants to Víctor Peinado and Javier Artiles.

References

1. Mandala, R., Tokunaga, T., Tanaka, H.: Combining multiple evidence from different types of thesaurus for query expansion. In: SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA, ACM (1999) 191–197
2. Smeaton, A.F., Quigley, I.: Experiments on using semantic distances between words in image caption retrieval. In Frei, H.P., Harman, D., Schäuble, P., Wilkinson, R., eds.: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'96, August 18-22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum), ACM (1996) 174–180
3. López-Ostenero, F., Gonzalo, J., Peñas, A., Verdejo, F.: Noun phrase translations for Cross-Language Document Selection. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001, Revised Papers. Volume 2406 of Lecture Notes in Computer Science., Springer (2002) 320–331
4. López-Ostenero, F., Gonzalo, J., Peñas, A., Verdejo, F.: Interactive Cross-Language Searching: phrases are better than terms for query formulation and refinement. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: Advances in Cross-Language Information Retrieval, CLEF 2002. Volume 2785 of Lecture Notes in Computer Science., Springer (2003)
5. López-Ostenero, F., Gonzalo, J., Verdejo, F.: UNED at iCLEF 2003: Searching Cross-Language Summaries. In: Evaluation of Cross-Language Information Systems, CLEF 2003. Volume 3237 of Lecture Notes in Computer Science., Springer (2004)

6. López-Ostenero, F.: Un sistema interactivo para la búsqueda de información en idiomas desconocidos por el usuario. PhD thesis, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (2002)
7. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: WordNet: An on-line lexical database. International Journal of Lexicography 3(4) (1990)
8. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet (1998)
9. Pirkola, A.: The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-Language Information Retrieval. In: Proceedings of SIGIR'98, 21st ACM International Conference on Research and Development in Information Retrieval. (1998) 55–63
10. Callan, J.P., Croft, W.B., Harding, S.M.: The Inquery Retrieval System. In: Proceedings of the Third International Conference on Database and Expert Systems Applications, Springer-Verlag (1992) 78–83