UNED@ImageCLEF 2004: using image captions structure and noun phrase based query expansion for Cross-Language image caption Retrieval.

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

NLP Group^{*}- Dept. de Lenguajes y Sistemas Informáticos, UNED c/ Juan del Rosal, 16 - E-28040 Madrid - Spain victor@lsi.uned.es, javart@bec.uned.es, {flopez, julio, felisa}@lsi.uned.es

Abstract

In this paper, we present the experiments for the first participation of the UNED's NLP Group in the ImageCLEF bilingual ad hoc task. Two different strategies are attempted: a) Expanding and translating queries with noun phrases, and b) Performing structured searches using entities located in topic titles over image caption fields.

All our experiments obtain results above the average MAP for the Spanish task. Our best result (using only structured searches over image captions) obtains 88.18% regarding the monolingual experiment and behaves a 8.3% better than Pirkola's structured queries. On the other hand, noun phrases expansion, which had proved successful in interactive tasks, influences negatively the results, due to the mismatch between the training corpus and the ImageCLEF collection.

1 Introduction

For its first participation in the ImageCLEF campaign, the UNED NLP Group decided to take 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 a list of topic statements and a collection of images with semistructured 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 nonrelevant image for the search). The collection was made up of 28,133 photographs from one of the most important sets of historic photography in Scotland.¹ All images had 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). Although not all fields contained text, this information was used in order to retrieve the relevant images.

The UNED team attempted two different strategies:

1. Expand queries with noun phrases. We expanded the queries using a bilingual Spanish-English phrase-based dictionary whose phrases had been automatically extracted from the CLEF news comparable corpus.

^{*}NLP Group website available at http://nlp.uned.es

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

- 2. Perform structured searches using the named entities and dates over the image caption fields, so that:
 - a) Proper names automatically detected in the query are searched in the "author" and/or "location" fields. If the search is non-nil, the search engine will favor images containing these entities tagged as author or location within the caption.
 - b) Temporal references and numbers in the query are searched in the same fashion on "date" fields. The latter were used on the assumption that they may be understood as years.

Two ideas, based on previous works, made us choose these strategies:

- 1. Expansion with Natural Language techniques is more beneficial with short queries [6] or short documents such as image captions [9]. Thus, we experiment with our query expansion technique based on aligned noun phrases, that gave excellent results in the interactive track (see [3, 4, 5]). Firstly, we test different approaches for batch Cross-Language Information Retrieval (CLIR) in order to evaluate the usefulness of our linguistics resources and then, we use the learned lessons to perform CLIR using the captions accompanying every image as short documents.
- 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 firstly discuss the possibilities of using noun phrases in query expansion and interactive tasks, we present the linguistic resources used (section 2.1), the preliminary CLIR experiments we performed previously over the CLEF collection (section 2.2) and we explain the settings of the ImageCLEF experiments (section 2.3). Below, we explain the procedure of structured searches over image captions used in this work (section 3) and comment the results of our official experiments (section 4). Finally, in section 5, we draw some conclusions and propose future work.

2 Query expansion with noun phrases

2.1 Linguistic resources

We used two comparable corpora, the Spanish and English CLEF corpora made up of the Spanish collection from the EFE news agency and the English collection from the Los Angeles Times newspaper, both of them from 1994.

We also had a bilingual dictionary containing a set of more than five million aligned noun phrases² which had been automatically recognized and extracted from the CLEF corpora using statistical data such as the frequency of sequences of two or three informative words (nouns and adjectives) in both languages [2]. We consider that two phrases are aligned if they have the same amount of informative words and there is one-to-one correspondence using a bilingual dictionary (e.g. see Table 1).

Previous UNED participations in the iCLEF tracks³ showed how helpful noun phrases could be in document selection [3], query translation and refinement ([4, 5]). Indeed, these phrases had proved successful in various tasks as intermediate translation units halfway between the word by

 $^{^2 \}rm Roughly,$ there are more than 4.5 million phrases containing two informative words and 850,000 containing three.

 $^{^{3}{\}rm The}$ Interactive track for the Cross-Language Evaluation Forum webpage is available at <code>http://nlp.uned.es/iCLEF</code>.

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: Sets of aligned phrases.

word translation (which can be made using a simple bilingual dictionary) and the translation of more complex structures such as sentences (often performed using a machine translation system).

Finally, we had a new bilingual Spanish-English dictionary made up from heterogeneous lexicographic resources such as usual dictionaries (some of them freely distributed), word lists and semantic networks such as WordNet [7] and EuroWordNet [10]. Every source went through an analysis and cleaning process in order to regularize their format and merge them, building up an XML-structured dictionary showing up all the information from each original source, without any 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 Mb⁴.

2.2 Preliminary experiments over the CLEF collection

In order to check the usefulness of noun phrases for a Cross-Language batch retrieval, firstly, we proposed the three following baselines based on the translation word by word using bilingual dictionaries:

- **naive baseline** Word by word translation, building a bag of words with all the possible translations appearing in our dictionaries.
- **frequencies** In this experiment, we built a bag of words from only those possible translations appearing in more than one lexicographic source, assuming that those ones which had high frequencies in our dictionary were 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 [8] and built the translated queries in a structured fashion, using the synonymy operators implemented in the INQUERY search engine [1] which permit to weight equally different possible translations for the same sense.

Then, we decided to compare these baselines with more sophisticated combinations of the available resources, such as:

- systran The queries were translated using the Systran machine translation system.
- **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 noun phrases containing it with the highest frequencies in the CLEF collection and then we translated them using their appropriate aligned phrases. 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 made use only of those phrases which contain at least two query terms.

 $^{^4\}text{Besides},$ when using this merged dictionary instead of the original VOX dictionary that we used in previous approaches, we obtain improvements in our CLIR experiments of about 36%

phrases + pirkola + systran As a last experiment, we performed one run combining the three resources, i.e. noun phrases, structured translations from our dictionaires and the Systran machine translation system.

As shown in Table 2, our experimental proposals using the noun phrases outperformed the baselines. The differences were statistically significant, according to a non-parametric Wilcoxon sign test. Our strong baseline even obtained the same average precision than Systran's translations, showing that the combination of Pirkola's structured queries and reliable lexicographic resources are good enough when searching over a collection of short documents.

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 final lessons we concluded from these experiments were the following:

- Phrases do improve CLIR results. Even though the percentage gain is not very high, in a setting with very small documents (e.g. image captions or topic titles) the difference could be higher.
- 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.
- The quality of a translation strongly depends on the resources used.

2.3 Settings for ImageCLEF experiments

Extending the above results to the ImageCLEF 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 system⁵ and the set of aligned noun phrases between English and Spanish.

Topic title were manipulated as usual in Cross-Language Information Retrieval tasks. Punctuation and stopwords were removed⁶ and informative words were lemmatized before attempting the translation using the dictionary or the expansion by means of the noun phrases.

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

3 Structured search using image caption fields

3.1 Entities recognizer

Before performing structured searches over the caption fields, we needed to locate the entities appearing in each topic title. So, we used our own recognizer which is able to find:

⁵Systran 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)...

- **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) ...)

3.2 Procedure of structured search over caption fields

After having located these entities, we performed the following procedure to identify the role of every entity in the ImageCLEF collection and to structure the query using the information appearing in the 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, firstly in Spanish and secondly 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 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. *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 On one hand, Roma, Irlanda, Norte, British Columbia, Canadá, Egipto, Londres, Bute, Escocia and York were correctly identified. On the other, St. Andrews, Cambridge, Tay Bridge, Crail Camp, North Street and Edimburgo was not.

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

Thus, this strategy strongly depends on the quality of the manual tagging of the collection. Sometimes it is not possible to identify the role of entities such as *Tay Bridge* or *St. Andrews* because they are not tagged this way in the image captions.

⁷We 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 experiences achieved after having performed the approaches explained on section 2.2, we decided to use the following strategies in our ImageCLEF experiments:

• Naive baseline using a word by word translation (UNEDESBASE).

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( tournament at st ) #phrase( final of the tournament )
#phrase( tournament at st ) #phrase( bond gains )
#phrase( title of chief )
```

To sum up, the set of submitted runs and its features are shown on table 4.

run	word translation	noun phrases	structured caption
UNEDESBASE	bag of words	Х	Х
UNEDES	Pirkola	Х	Х
UNEDESENT	Pirkola	Х	\checkmark
UNEDESENTNOO	Pirkola	\checkmark	\checkmark
UNEDORENTNOO	Pirkola		\checkmark

Table 4: UNED's set of submitted runs.

4.2 Results

Our five experiments obtained very promising results,⁸ and they all are placed above the average (notice that average MAP for Spanish is 0.30). The best one, UNEDESENT, has even gotten the second best result in the ranking among all the participants in the Spanish-English task, with losses of only 11.82% and 0.77% with regard to the monolingual experiment and the Sheffield best run,⁹ respectively, as shown in Table 5.

Structured queries over image captions (UNEDESENT) obtained an improvement of about 8.3 % with regard to the Pirkola's proposal (UNEDES). It seems to indicate that such a simple detection

⁸Official results are available at http://ir.shef.ac.uk/imageclef2004/adhoc_results/adhoc_results.html. ⁹Sheffield's es_es_fb has been the best run among all the Spanish participants.

of entities can improve the average results, taking advantage of the information contained in the image captions.

Unsurprisingly, expansion using noun phrases does not improve the structured search. Our set of aligned noun phrases had been previously extracted from a different collection, out of the photographic domain. As shown on section 4.1, the expansion inserted too much noise in the form of unrelated noun phrases, such as "gains after a bond", "chiefs into the title" ...

Finally, notice that our weakest baseline (UNEDESBASE) is about 26% better than the average MAP for the Spanish participants which seems to point the reliability of our dictionary.

run	\mathbf{MAP}	% monolingual
Sheffield	0.52	88.86
UNEDESENT	0.52	88.18
UNEDES	0.48	82.32
UNEDESENTNOO	0.47	79.66
UNEDORENTNOO	0.42	71.93
UNEDESBASE	0.38	64.27
average	0.30	50.87

	Table 5	5:	Results	for	UNED's	runs.
--	---------	----	---------	-----	--------	-------

5 Conclusions and future work

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.

The results obtained by using this strategy seems to indicate that noun phrases should be extrated from the appropriate corpus, in order to avoid noise.

• Perform structured search using entities located from topic titles over the image caption fields.

This straightforward refinement obtains an improvement of 8.3 % in relation to Pirkola's baseline and can be easily extended to other searches over structured documents.

• Given that our weakest baseline is above the average MAP, our basic linguistics resources have proved very useful.

As possible lines of future work, we are working on:

- Extracting the noun phrases directly from the appropriate collection. So far, we have been using the CLEF comparable corpora due to their size but this decision implies too much noise.
- Refining our recognizer in order to make a better detection of entities.
- Quitting using commercial machine translation systems which are out of our control.

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project R2D2 (TIC2003-07158-C0401), and a grant from the UNED (*Universidad Nacional de Educación a Distancia*).

References

- J. P. Callan, W. B. Croft, and S. M. Harding. The Inquery Retrieval System. In Proceedings of the Third International Conference on Database and Expert Systems Applications, pages 78–83. Springer-Verlag, 1992.
- [2] F. López-Ostenero. 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, December 2002.
- [3] F. López-Ostenero, J. Gonzalo, A. Peñas, and F. Verdejo. Noun phrase translations for Cross-Language Document Selection. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, 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, pages 320–331. Springer, 2002.
- [4] F. López-Ostenero, J. Gonzalo, A. Peñas, and F. Verdejo. Interactive Cross-Language Searching: phrases are better than terms for query formulation and refinement. In C. Peters, M. Braschler, J. Gonzalo, and M. Kluck, editors, Advances in Cross-Language Information Retrieval, CLEF 2002, volume 2785 of Lecture Notes in Computer Science. Springer, 2003.
- [5] F. López-Ostenero, J. Gonzalo, and F. Verdejo. 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] R. Mandala, T. Tokunaga, and H. Tanaka. 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, pages 191–197. ACM, 1999.
- [7] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4), 1990.
- [8] A. Pirkola. 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, pages 55–63, 1998.
- [9] A. F. Smeaton and I. Quigley. Experiments on using semantic distances between words in image caption retrieval. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, 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), pages 174–180. ACM, 1996.
- [10] P. Vossen. Introduction to EuroWordNet. Computers and the Humanities, Special Issue on EuroWordNet, 1998.