UNED at ImageCLEF 2005: Automatically Structured Queries with Named Entities over Metadata.

Víctor Peinado, Fernando López-Ostenero and Julio Gonzalo
NLP Group, ETSI Informática, UNED
c/ Juan del Rosal, 16, E-28040 Madrid, Spain
{*victor, flopez, julio*}*@lsi.uned.es*

Abstract

In this paper, we present our participation in the ImageCLEF 2005 ad-hoc track. First, we describe a preliminary pool of cross-language experiments with the ImageCLEF 2004 testbed performed in order to evaluate the impact of different-size dictionaries using three distinct approaches. These differences are not remarkable, however recognizing named entities and launching structured queries over the metadata improve the results in all cases. Then, we decided to refine our named entities recognizer and repeat the three approaches with the 2005 topics, achieving the best result among all cross-language European Spanish→English runs.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries

General Terms

structured queries

Keywords

image retrieval, cross-language information retrieval, named entities recognition

1 Introduction

In this paper, we describe the experiments submitted by UNED to the ImageCLEF 2005 ad-hoc task track.

On one hand, we explain a first pool of preliminary experiments using the ImageCLEF 2004 testbed and the Spanish official topics, performed in order to study the impact of different-size dictionaries in the final results. We attempted three different approaches: i) a naive baseline using a word by word translation of the title topics; ii) a strong baseline based on Pirkola's work [5]; and iii) a structured query using the named entities with field search operators and Pirkola's approach.

Our best runs achieved an average precision of .54, outperforming both our last year's participation and the best official cross-language run. The differences among dictionaries were not remarkable, except for the smallest one, which obtained lower precision values. However, we did

confirm interesting differences among approaches: runs based on structured queries were substantially better than the others.

On the other hand, we describe the UNED's participation in the ImageCLEF 2005 track. Given the benefits of recognizing named entities in the topics in order to structure the queries, we decided to improve our recognition process. Now we are able to locate and identify more complex proper nouns, temporal references and numbers. Then, we performed the three approaches over the 2005 testbed obtaining the first and second best cross-lingual runs in European Spanish, representing the 94% of our monolingual experiment.

The structure of this paper is the following: in Section 2, we explain the preliminary set of experiments using the ImageCLEF 2004 testbed and different-sized bilingual dictionaries. Then, in Section 3, we focus on our participation in ImageCLEF 2005. In section 3.1, we describe this year's settings, in section 3.3 the runs submitted to this year's edition of ImageCLEF are presented and then, in section 3.4, we comment the results obtained. Lastly, in Section 4, we draw some conclusions.

2 Preliminary experiments

2.1 Testbed

For our preliminary series of experiments, we used the Spanish-English ImageCLEF 2004 ad-hoc track testbed [2], which consisted of 25 topics written in Spanish, a set of 28,133 photographs annotated with rich semi-structured captions and a pool of relevance judgments generated at the track.

Every original topic in English consisted of a title and a more extensive narrative fragment describing an information need. We used the Spanish version of the topics, in which only titles were translated.

Images had an accompanying textual description consisting of eight human-annotated metadata fields, such as: a unique ID, both short and long titles, the location where the image had been taken, a description of the image, the date, the author's name and some general categories in which the photograph may be included (e.g. `[ferries]`, `[woods & forests]`, `[panoramic views]`, ...). Even though this information was not always complete, we decided to use it in order to improve our retrieval of relevant images.

All retrieval runs have been done with the Inquery search engine [1], which provided the rich query language required in our approach.

2.2 Locating and identifying named entities

In a first step, our simple recognizer uses a straightforward set of rules in order to locate all named entities appearing in the ImageCLEF topics and looking like such (see [3]):

- Expressions in uppercase wherever uppercase is not prescribed by punctuation rules are tagged as possible **proper nouns**.
- Expressions matching words such as weekdays, months or seasons are tagged as **temporal references**.
- Then, any numerical expression is tagged as a **number**.

Then, we attempt to classify these possible proper nouns and temporal references by checking if they appear as author's names, locations and dates in the collection. If they do, we build the query re-structuring this information so that our search engine favors those images with such metadata.

The procedure is the following:

1. If it is a proper noun, 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 is a numerical expression, 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. Finally, if it is a temporal reference, we check if it is a date in a similar fashion (see [3] for further details).

2.3 Experiments

We compared the three following approaches:

i) A **naive baseline** using a word by word translation. Words which were not present in our bilingual dictionary were left untranslated. For instance, the query for topic 2 (*Fotos de Roma que fueron tomadas en Abril de 1908*) was translated and built using Inquery's operators as:

```
#sum(pointless unpointed tomada taken taken take assume take take get espouse take on
take take capture seize have take imbibe take april 1908)
```

ii) **strong baseline** following Pirkola's proposal [5], where alternative translations for a query term were taken as synonyms, giving them equal weights:

```
#sum(#syn(pointless unpointed) tomada #syn(taken taken) #syn(take assume take take get
espouse take take capture seize have take imbibe take) april 1908)
```

iii) Our structured query approach, which incorporated **field search** operators in addition to Pirkola's strategy:

```
#field(DDATE #sum(1908)) #field(DDATE #sum(april)) #field(SOURCE #sum(rome))
#sum(#syn(pointless unpointed) tomada #syn(taken taken) #syn(take assume take take get
espouse take take capture seize have take imbibe take) april 1908)
```

We tried all three conditions with six different bilingual dictionaries: **FreeDict** a freely available *on line* dictionary; **EWN** generated from the official EuroWordNet multilingual semantic network; **EWN2** compiled from an updated version of the Spanish Wordnet; **Vox** an electronic version of the Vox-Harraps Spanish-English dictionary; [1] **All-Vox** a combination of all the dictionaries above except Vox; and finally, **All** a merged version of all four dictionaries.

Finally, we 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 Section 2.2) and an additional cross-language run where named entities and temporal references are annotated manually. The latter was intended to evaluate the effects of errors in the automatic location of entities.

2.4 Results and discussion

For all bilingual dictionaries, our structured query approach was better than the naive and Pirkola baselines. Pirkola's approach was, in turn, substantially better than its naive counterpart in all cases. However, only the differences between our structured query approach and the naive baselines were relevant according to a non-parametric Wilcoxon sign test (in half of the cases).

The differences among dictionaries were not statistically significant either. For most of the runs, the translations provided by EWN2, All-Vox or Vox separately were enough to reach the highest precision values.

[1] This is the only genuine bilingual dictionary that we used and we took it as the basis for a merged version of all dictionaries.

Our best runs achieved an average precision of .54, which represents 91% of our best monolingual run. This result slightly outperformed 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). For further details about the results, see also [4].

Remarkably, the manual annotation of named entities did 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.

3 ImageCLEF 2005 experiments

3.1 Settings

As in the previous edition, the testbed provided to ImageCLEF 2005 ad-hoc task participants was the St Andrews image collection. In this case, the participants were given 28 topics, each containing a title and a narrative fragment with verbose details about an information need.

Besides, this year there have been proposed two distinct set of Spanish topics which tried to show the local variants of the language: one European Spanish translation and another Latin American version. Even though the topics had been translated into Spanish wholly, we only took the short titles in our experiments.

3.2 Entities, temporal references and numbers found

Regarding our last year's experience, some improvements have been done in our named entities recognition process. Now, we can locate more complex multi-word proper nouns and temporal references by attaching several simple entities of the same type usually connected by articles, prepositions and conjunctions. And so, our recognizer is able to locate some Spanish named entities such as the ones shown in Table 1.

In Table 2, we show the named entities located in the ImageCLEF 2005 European Spanish topics.

It is worth mentioning that topics proposed contained fewer expressions likely to be named entities than last year. Indeed, no temporal reference or number was located and we only could take advantage of the improvements of the recognizer in 6 out of 28 topics. Regarding the precision of the recognition, notice that the entities located in this year's topics are the same that a user would have manually selected.

3.3 Submitted runs

We submitted to ImageCLEF 2005 track five different runs, based on the same runs we had already tested in Section 2.3. First, one monolingual run in order to establish the maximum precision that we could achieve using our resources. Then, a naive run building the queries with a simple word by word translation. We also submitted two runs based on the **strong baseline** with the synonymy's operators which allowed us to enrich and expand the translations while minimizing the noise. Lastly, we repeated the run adding the field search operators.

The features that define each of these runs are shown in Table 3.

3.4 Results and discussions

The official results obtained by our five runs are shown in Table 4. First of all, it is worth mentioning that our cross-lingual run enriched with named entities `unedESENEnt` obtained the best MAP score among all official cross-lingual runs having European Spanish as the source language. Its counterparts without using the named entities `unedESEN` and `unedESAmerEN` got comparable results: .28 (3^{rd} position in European Spanish) and .26, respectively. On the other hand, our simpler cross-lingual run achieved .19.

organizations
Alta Comisaría de las Naciones Unidas para los Refugiados
Consejo de Seguridad de la ONU
Orquesta Sinfónica de la Radio Bávara
Comisión Nacional del Mercado de Valores
Agencia Internacional de la Energía Atómica
Fuentes del Centro de Coordinación de Salvamento Marítimo de Galicia
temporal references and dates
ocho de la tarde de ayer 31 de diciembre
domingo 2 de enero de 1994
noche del miércoles
medianoche del 31 de diciembre
16,30 de ayer viernes
mediodía de ayer domingo
3 ene
noche del 20 de noviembre de 1992
septiembre y octubre de 1993
2 de octubre y 8 de diciembre
octubre de 1991 y noviembre de 1992
cardinal numbers
siete millones
20.000 millones
treinta y cuatro
mil novecientos
tres mil millones
ochocientos sesenta millones
95,500 y 95,750

Table 1: Examples of Spanish named entities: proper nouns and organizations, temporal references and cardinal numbers located in the EFE news agency corpus.

In spite of the apparently poor result obtained by our monolingual run, it is remarkable the small difference regarding our best cross-lingual run, whose MAP score represents 94% of `unedmono`'s. This leads `unedESENEnt` even closer than our last year's best strategy.

4 Conclusions

In this paper, we have presented our participation in the ImageCLEF 2005 ad-hoc track.

First, we have described a preliminary pool of cross-language experiments with the ImageCLEF 2004 testbed performed in order to evaluate the impact of different-size dictionaries using three distinct approaches. We outperformed our ImageCLEF 2004 participation but the differences among dictionaries were not remarkable. However, in all cases, our results dramatically improved when recognizing named entities and launching structured queries over the metadata. So, we decided to refine our named entities recognizer and repeated the three approaches with the 2005 topics, achieving the best result among all cross-language European Spanish→English runs.

Therefore, automatic query structuring seems an effective strategy to improve cross-language retrieval on semi-structured texts. Remarkably, no sophisticated named entity recognition machinery is required to benefit from query structuring. Of course, it remains to be checked whether this result holds on collections with different metadata fields and different textual properties.

topic #	Entities identified
1	avión en tierra
2	gente reunida en torno a un quiosco de música
3	perro sentado
4	barco de vapor atracado
5	estatuas de animales
6	pequeño barco de vela
7	pescadores en un barco
8	edificio cubierto de nieve
9	caballo tirando de un carro o carruaje
10	**imágenes del** [$_{PN}$ **Sol**], [$_{PN}$ **Escocia**]
11	paisaje de montañas suizas
12	**postales de** [$_{PN}$ **Iona**], [$_{PN}$ **Escocia**]
13	viaducto de piedras con arcos
14	gente en el mercado
15	tiro al hoyo en el green
16	olas rompiendo en la playa
17	hombres o mujeres leyendo
18	mujer con vestido blanco
19	**postales compuestas con imágenes de** [$_{PN}$ **Irlanda del Norte**]
20	**visita real a** [$_{PN}$ **Escocia**] **(excepto a** [$_{PN}$ **Fife**]**)**
21	**monumento al poeta** [$_{PN}$ **Robert Burns**]
22	edificio con bandera al viento
23	tumba en el interior de una iglesia o catedral
24	primer plano de un pájaro
25	puerta en forma de arco
26	retratos de grupo de personas de ambos sexos
27	mujer o niña llevando una cesta
28	**fotografías a color de bosques alrededor de** [$_{PN}$ **St. Andrews**]

Table 2: Named entities identified for each topic title.

Acknowledgments

This work has been partially supported by the Spanish Government under project R2D2-Syembra (TIC2003-07158-C04-02). Víctor Peinado holds a PhD grant by UNED (*Universidad Nacional de Educación a Distancia*).

References

[1] 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] P. Clough, M. Sanderson, and H. Müller. The CLEF Cross Language Retrieval Task (ImageCLEF) 2004. In *Cross Language Evaluation Forum, Working Notes for the CLEF 2004 Workshop*, volume 3491 of *Lecture Notes in Computer Science*. Springer Verlag, 2005.

[3] V. Peinado, J. Artiles, F. López-Ostenero, J. Gonzalo, and F. Verdejo. UNED at Image CLEF 2004: Detecting Named Entities and Noun Phrases for Automatic Query Expansion and Structuring. In *Cross Language Evaluation Forum, Working Notes for the CLEF 2004 Workshop*, volume 3491 of *Lecture Notes in Computer Science*. Springer Verlag, 2005.

Dimensions	unedmono	unedESENnaive	unedESEN	unedESamerEN	unedESENent
Query language	EN	ES (*Eur.*)	ES (*Eur.*)	ES (*Amer.*)	ES (*Eur.*)
Initial query	title	title	title	title	title
Query type	automatic	automatic	automatic	automatic	automatic
Feedback/expansion	X	√	√	√	√
Modality	text	text	text	text	text

Table 3: Dimensions defining the runs.

run	**MAP**	**variation**
unedmono	.34	–
unedESENEnt	.32	94%
unedESEN	.28	82%
unedESAmerEN	.26	76%
unedESENnaive	.19	56%

Table 4: Results of our official runs

[4] V. Peinado, F. López-Ostenero, J. Gonzalo, and F. Verdejo. Searching Cross-Language Metadata with Automatically Stuctured Queries. In *European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005)*, volume 3652 of *Lecture Notes in Computer Science*. Springer Verlag, September 2005.

[5] 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.