Web Track for CLEF2005 at ALICANTE UNIVERSITY

Trinitario Martínez, Elisa Noguera, Rafael Muñoz, Fernando Llopis
Department of Software and Computing Systems
University of Alicante, Spain
{tme, elisa, rafael, llopis}@dlsi.ua.es

Abstract

This paper presents the first experiment done for the CLEF2005 Multilingual Web Track. At present conference we have focused our main effort in the Spanish part of the Mixed Monolingual task, but we have also participated in others several languages and in the Bilingual English-Spanish task. A passage based IR system is applied at retrieving phase. Also a language identifier has been created in order to build a full automatic system without the need of knowing the topic language.

Categories and Subject Descriptors
Information Search and Retrieval

General Terms
Measurement, Experimentation

Keywords
Information retrieval, question answering

1 Introduction

The Cross Language Multilingual Web Retrieval (WebCLEF) track consists of the evaluation of Information Retrieval systems on noisy multilingual documents. Particularly, the WebCLEF document collection consists of webpages from European governmental sites for at least 10 languages/countries.

Retrieving in a Multi/Crosslingual manner is a natural and common established way for carrying out web searches. The aim of this specific task is to find the correct document on which the topic description is. This paper is structured as follows: next section describes the collection and topics used, later we explain the corpora processing and retrieving. Afterwards we show the results and conclusions, and finally we make discuss about future improvements of the system.

2 Processing phase

2.1 Data Specifications

The targeted corpus is a mix of governmental sites in Europe. More concretely, the collection, EuroGOV, consists of web documents crawled from European governmental sites. Here's a list of (top level) domains from which pages are included:

at, be, cy, cz, de, dk, ee, es, fi, fr, gr, hu, ie, int, it, lt, lu, lv, mt, nl, pl, pt, ru, se, si, sk, uk

The amount of data is impressive: over 20 gigas of compressed text files containing diverse governmental information on multiple types, such us HTML, ZIP, DOC and PDF format. Documents are gathered in a pseudo-XML format, storing domain, url, id, md5 signature, type (html, doc, pdf…) and data (in binary or text format). This corpus has been very controversial, and finally just html documents were designed to be retrieved by organizers.

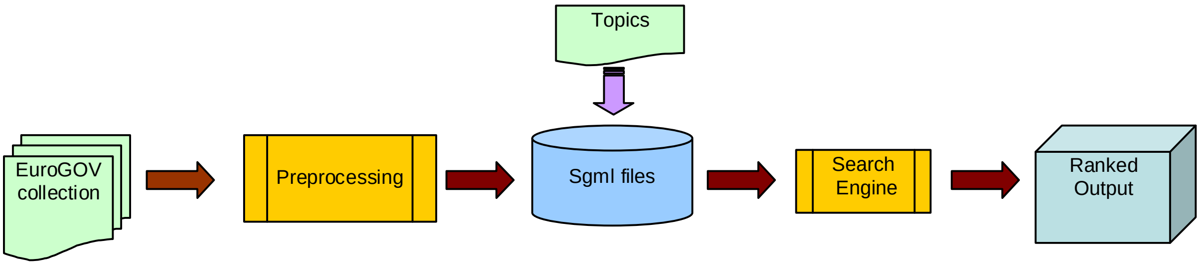
2.2 Data Preprocessing

At our first participation in this kind of competitions, we have focused our efforts in Spanish monolingual queries, and have made some others symbolic approaches. We have divided the corpus by language. This is required in order to not managing the whole amount of data.

Once html files are extracted from the corpus:

1. Firstly, META labels are collected from the files. Specifically, *title* and *keywords* labels are saved for the retrieval phase.
2. Second step consists of replacing HTML code by its equivalent, as for example “&raquo;” ó “>”.
3. Thirdly, regular expressions are used in order to remove special tags, obtaining a plane text.
4. At the end of the process, id, keywords, title and plane text of each document are stored in sgml files in order to conform a correct input for the Information Retrieval system (Trec format).

WebCLEF Processing



We also have developed a language identifier with the purpose of fully automating the Mixed Monolingual process.
In addition to this, we had built up one specific module to extract pdf, doc and zip files from EuroGOV, but this has not been used because organizers decided do not retrieving this files types.

2.3 Topic creation

As this has been the first Multilingual Retrieval Track at CLEF, topics have been developed by participants.

Queries are based on a collection of 547 multilingual topics. These are classified in two categories:

- Ø Home Page finding: a homepage web is searched (i.e. www.dlsi.ua.es).
- Ø Named Page finding: a specific non-homepage is searched in this case (i.e. http://www.dlsi.ua.es/cgi-bin/wwwadm/personal.cgi?id=eng&nom=rafael&tipus=pdi).

At this phase, we created 30 monolingual known-item topics (15 named-page and 15 home page topics) in Spanish.
We also detected identical or similar pages in the collection by the use of search engines, and also by manual searches through the corpuss in order to produce consistent and well-formed topics. Also an English translation of the topic statement is provided with the purpose of being used in the multilingual task. For example, if we have a topic with this title:

Presidente del gobierno

In the traduction, the *Spanish* adjective is added to make more precise the future search through the whole corpuss:

Spanish government president

We developed several topics with .PDF and .DOC files which were finally discarded by organizers because of some participants found some problems with these formats at extraction text task.

2.4 Retrieving phase: IR-n system

IR-n is a passage retrieval system (RP). RP systems [6] locate in contiguous fragment of text (passages) and boost IR field by proposing a set of solutions to tradicional IR systems common problems. One of the main advantages of these systems is that they allow us to determine not only if a document is relevant or not, but also the detection of the relevant part of the document.

IR-n system uses the sentences as atoms with the aim to define the passages. The passages are usually composed of a fixed number of sentences. This number has a great dependency of the targeted collection. Furthermore, IR-n system uses overlapping passages in order to avoid that some documents cannot be considered relevant if words of the question appear in adjacent passages.

For every language, the resources used were provided by the CLEF organizers (http://www.unine.ch/info/clef). There are stemmers and stopword lists for all languages, with the lack of Danish and Dutch stemmers.

IR-n system allows the use of distinct similarity measures (Ex. Okapi [7]). This involves an advantage, so that, in each task is used the best similarity measure.

With the aim of being able to indexing the documents in html format, indexing module has been modified to consider the tags *title* and *keywords.* The words which are in these labels have more weight than the words of the rest of the document in order to increase the value of the documents which have words of the query in the labels than the rest one.

According to others IR systems, IR-n system uses different techniques of the query expansion. Previous researches [8] have showed that the approaches get better results when they are based on passages and in the complete document.

Finally, this year for the adhoc task has been implemented a technique called combined passages [9]. It applies fusion methods, which are used in multilingual tasks to combine results with the different size of passages.

3 WebCLEF Tasks

Although we have focused our brief in the Spanish competition, others languages have been taken into account. The targeted languages have been:

Mixed Monolingual task:

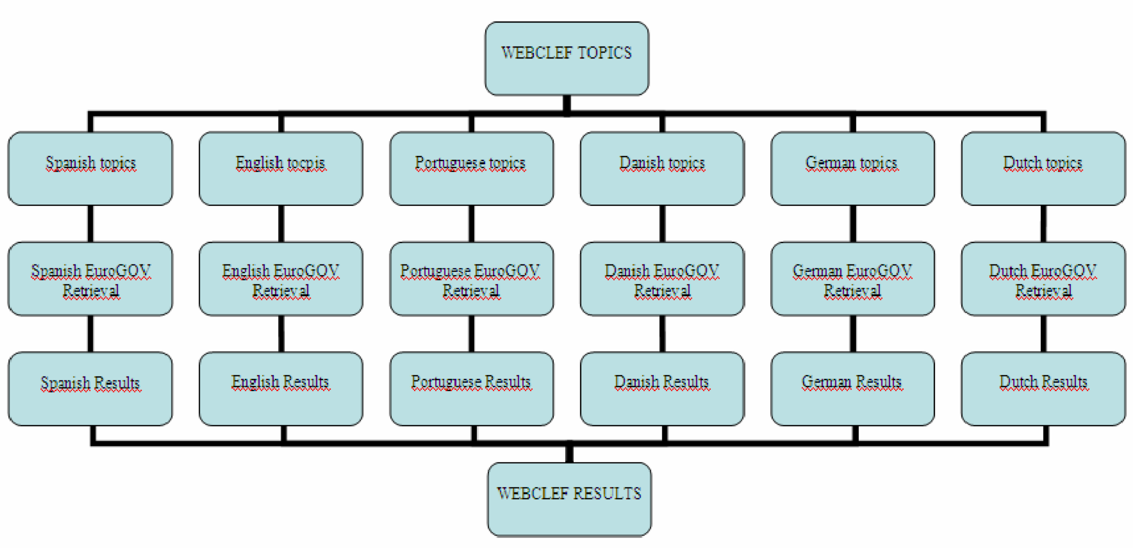
- Danish
- Spanish
- Dutch
- German
- English
- Portuguese

Bilingual task:

- English - Spanish

3.1 Mixed Monolingual task

For the monolingual task, topics have been divided by language so that they are individually processed by the system. The specific results are finally merged in a results file.



Note that other languages topics, like Hungarian, Polish, French, Greek, Icelandic and Russian where not taken into account because we have not resources of these languages.

3.1.1 Language identification

As a baseline run, we have developed a language detector in order to automatically distinguish the correct language of the topic. In particular, our language detector has this general bases:

- Dictionary based (joined dictionaries, specific per-language stopwords)
- Characterised part-of-word terminology (i.e. “ção” in the case of Portuguese)
- Specific governmental terminology (i.e. “administration” in the case of English)

This philosophy gave us a good response in Spanish, English, Portuguese and Danish. Lamentably, Dutch and German are too much similar, and the system becomes occasionally erroneous. We have not reliable experience with these languages.

Once language topics were identified, they were separately stored in different files and run with the specific part of the EuroGOV corpus. By this way, a faster response of the system is obtained than when whole corpus is taken.

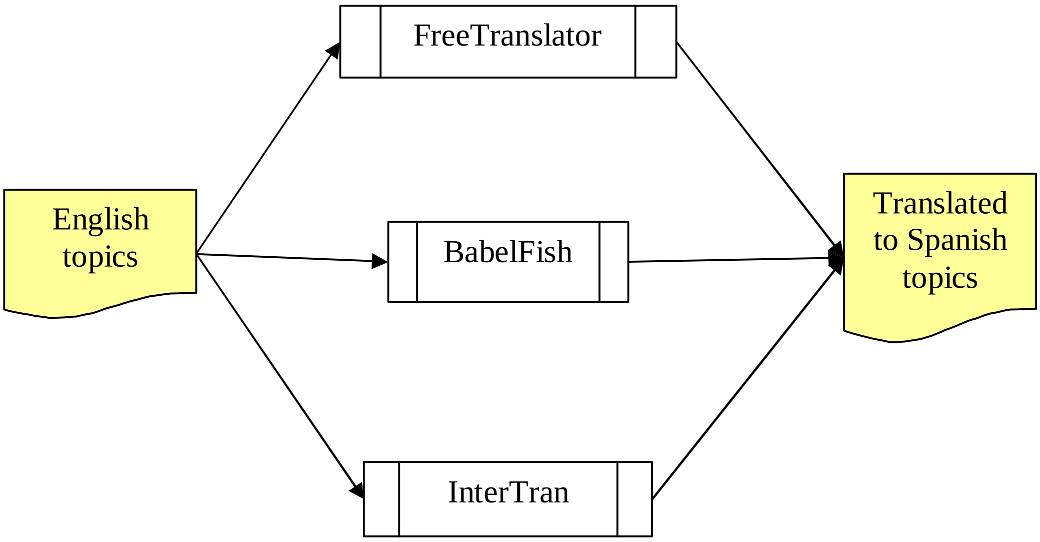
Table 1: Language identification stadistics

Language	# Corrects	# Incorrects
Spanish	134	2
English	117	2
Portuguese	56	0
Dutch	52	15
German	44	2
Danish	29	1

As statistical results, just to mention that the language identificator could not was capable of determinate the language of seven topics from the Spanish, English, Portuguese, Dutch, German and Danish set. The rest of languages (87 topics) were not taken into account because they were not later processed by the IR system.

3.2 BiEnEs task

The BiEnEs (Bilingual English-Spanish) task consists of carrying out searches in the Spanish corpuss of EuroGOV from topics written in English. Our automatic approach has been performed by a merging of three different on-line translators. The main idea is that the more common word is, the higher relevancy has.
The used translators have been Freetranslator[1], BabelFish[2] and InterTran[3]. An example of this is shown in next picture:



[1] http://www.freetranslation.com/
[2] http://world.altavista.com/
[3] http://www.tranexp.com/win/itserver.htm

4. Results

4.1 Monolingual task results

In the process of our first experiment at WEBClef2005, we have focused on the Spanish Topics part of the Mixed Monolingual task. Also to mention Spanish Topics is the greater subset of the topic set. So, this is an important part of the task. We also have been doing experiments with other five languages: Danish, German, English, Dutch and Portuguese. On next table, averages at 1, 5, 10, 20 and 50 are shown, as the MRR too. The last column shows the difference between our system and the average results.

Table 2: Mixed Monolingual official results per language

Language	Aver. At 1	Aver. at 5	Aver. at 10	Aver. at 20	Aver. at 50	MRR	Dif
ES	0.1716	0.3134	0.3433	0 .3731	0.4328	0.2377	+4,4261
DA	0.0333	0.0667	0.0667	0.0667	0.0667	0.0500	-4,082
DE	0.1579	0.2105	0.2632	0.3158	0.3158	0.1907	-9,4245
EN	0.0496	0.0744	0.0826	0.0826	0.0909	0.0614	-15,2636
NL	0.1356	0.1525	0.1525	0.1695	0.1695	0.1451	-9,4245
PT	0.0508	0.1695	0.1695	0.2034	0.2712	0.0833	-6,2003

On next table, results of the application of the automatic language detection at the Mixed Monolingual task are shown. Obviously, results are lower than previous, and give us an idea about how a mechanized system would response. Accidentally, one erroneous topic numeration run was submitted, but later another run was made. Finally, results are shown:

Table 3: Mixed Monolingual with automatic language detection results per language

Language	Aver. at 1	Aver. at 5	Aver. at 10	Aver. at 20	Aver. at 50	MRR
ES	0.1343	0.2612	0.3134	0.3582	0.4104	0.1995
DA	0.0333	0.0667	0.0667	0.0667	0.0667	0.0500
DE	0.0702	0.1053	0.1579	0.2105	0.2105	0.0942
EN	0.0496	0.0744	0.0826	0.0826	0.0909	0.0614
NL	0.0847	0.1017	0.1017	0.1186	0.1186	0.0943
PT	0.0508	0.0847	0.1017	0.1525	0.2203	0.0656

4.2 Bilingual English-Spanish results

Clearly, results obtained at this task are influenced by the results of the Spanish Monolingual task and also by the association of the three mentioned translators.

Table 4: Bilingual English-Spanish task

Aver. at 1	Aver. at 5	Aver. at 10	Aver. at 20	Aver. at 50	MRR	Dif
0.0299	0.0522	0.0597	0.0746	0.0970	0.0395	-2,5028

5 Conclusions

In this paper we have presented the first version of our system at the Multilingual Web Track at CLEF. We have targeted the Mixed Monolingual Task, concretely Spanish, Danish, Dutch, German, English and Portuguese languages. At Spanish, we are above the average, whilst at other languages the system has a lower performance (we have never worked before with Danish nor Dutch). More time would be desirable in order to finish the whole system, and tuning it.

At the automatic language detection process, we lack of the need of a better language detector. The one used here has been a fast developed attempt, but not perfect.

At the Bilingual English to Spanish task, the conclusion is clear: general purpose translator is not a good tool to be used here, due to the fact that the retrieving collection is focused in a determined scope like governmental processes are. Our 3-translator association works better than one translator in its own, but this is not the ideal solution, and we consider the requirement of a specialized translator a must.

Finally, sometimes we have found that Keywords tags extracted from EuroGOV corpuss were adding noise to the system, because some HTML document can have several governmental scope keywords. This is why they are not working perfectly and getting in worse results.

6 Future works

A way to improve our proposed system in future would be to extend our Mixed Monolingual task in order to include missing languages at this participation (Hungarian, Polish, French, Greek, Icelandic and Russian). Our major lack here is the necessity of resources (stemmers, stopwords lists and so on).

Another good advance would be to experiment with hyperlinks of the HTML documents of EuroGOV Collection, storing them and establishing some kind of relation between web pages. Also a little extraction of the link text string can add more information to retrieve.

A way to progress in automatic processing with language identification phase would be improving the present identifier in the way it could use n-grams, and some discriminatory and specific EuroGOV corpuss machine learning language acquisition would be performed.

We aim to extend the system so that Multilingual task could be fully run on next WebCLEF participation. This will require the extraction of language cues by a specific ad-hoc detector.

Acknowlegments

This work has been partially supported by the Spanish Goverment (CiCYT) with grant TIC2003-07158-C04-01 and also by the Regional Tecnology Ministry of Valencia Government by means of the projects with reference GV04B-276 and GV04B-286.

References

[1] Llopis, F., Muñoz, R, Noguera, E., M. Terol, R. IR-n r2: Using normalized passages. CLEF 2004

[2] Callan, J. P.: Pasaje-Level Evidence in Document Retrieval. In Proceedings of the 17th Annual Internacional Conference on Research and Development in Information Retrieval, London, UK. Springer Verlag (1994) 302-310.

[3] WebCLEF. Cross-lingual web retrieval, 2005. http://ilps.science.uva.nl/webclef/

[4] WinEdt Dictionaries. http://www.winedt.org/Dict/

[5] Rafael M. Terol, Patricio Martínez-Barco, Fernando Llopis, Trinitario Martínez: An Application of NLP Rules to Spoken Document Segmentation Task. NLDB 2005: 376-379

[6] M. Kaskziel and J. Zobel. Passage retrieval revisited. In *Proceedings of the 20th annual International ACM Philadelphia SIGIR*, pages 178–185, 1997.

[7] Aitao Chen and Fredric C. Gey. Combining query translation and document translation in cross-language retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, pages 108–121, Trondheim, Norway, 2003. Springer-Verlag.

[8] Aitao Chen and Fredric C. Gey. Combining Query Translation and Document Translation in Cross-Language Retrieval. 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003.108-121.

[9] Llopis F., Noguera E. Combining passages in monolingual experiments. In *Workshop of Cross-Language Evaluation Forum (CLEF 2005)*, In this volume, Vienna, Austria, 2005.