UNED at WebCLEF 2005

Javier Artiles, Víctor Peinado, Anselmo Peñas, Felisa Verdejo
Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
c/ Juan del Rosal, 16, Ciudad Universitaria, 28040 Madrid - Spain

Abstract

This paper describes UNED NLP & IR Group experiments at the Web CLEF track. The UNED NLP & IR Group took part in the bilingual English-Spanish task. Our main attempt is to improve query translation by means of a non supervised approach to the translation of out-of-vocabulary words. Indexing is made preserving the information about the most relevant structural elements (title, metadata, headings, etc.). Search process is also explained and, finally, experimental results are presented.

Categories and Subject Descriptors

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

General Terms

Measurement, Performance, Experimentation.

Keywords

Information Retrieval, Cross-Language Information Retrieval

1 Introduction

For the first participation in the Web CLEF track, the UNED NLP & IR Group took part in the bilingual English-Spanish task. The main goal of the track is to test information retrieval systems in a cross-language environment settled at the Web. In all the proposed tasks there was a set of known item topics and a collection of web pages. The participant's systems had to find a particular page for each search topic.

The WebCLEF organization proposed to the participants an evaluation where the systems must return a ranked list of results (50 hits maximum) for each known item topic. The retrieved results should contain the target web page, in the first position of the ranking or as near as possible. In the case of the bilingual task, the set of 134 topics in English targets pages in Spanish that can be found at the Eurogov collection [1].

2 Indexing

The EuroGov collection was indexed with Lucene 1.4.3. (http://lucene.apache.org). Each HTML document was reduced to plain text by removing the HTML tags. Given the wide range of languages at the collection, we decide to discard word normalization techniques and stopwords lists, both in documents and queries. We extracted from the HTML structure the following fields:

title, metadata, headings (text tagged with h1, h2, etc), body, outgoing links and it's corresponding text anchors. The text present at each field was tokenized using Lucene's *StandardAnalyzer* and assigned to a document identified in the index by its URL and it's EuroGov ID.

The EuroGov collection is composed of 3,589,501 web pages from the European governmental sites (including 27 different web domains, as *.uk, .gov, .fr, .es*, etc.). The target Spanish page of a query could be hosted at any of these domains. In order to work with a smaller dataset we decided to index only the Spanish pages, according to the language detection output provided by the organization.

3 Query Translation

The approach to the query translation has been synthesized in a very simple algorithm that performs a word-by-word translation. We translate the queries word-by-word using a bilingual dictionary. If the word can't be found in the dictionary we proceed to extract a candidate translation from the Web by means of an algorithm explained in this section.

In first place we have to decide which words of the query need to be translated to Spanish or must remain untranslated. Usually this step is implied by the use of a dictionary. Word out of the vocabulary of the dictionary are considered not translatable. In our case, in addition to the dictionary, we use a method to translate words out of the vocabulary. That's because it is important to decide if it is really necessary to translate a word. To confront this problem, we profited from the language detection output of the EuroGov collection provided by the Organization and we created separated collections by language. Our hypothesis is that a word which has its highest df/N (where df is the document frequency of the word and N the number of documents at the collection) in a collection of documents other than the English Collection is not an English word, and hence must not be translated. Furthermore, in that cases it is likely to find that the highest df/N corresponds to the Spanish Collection, given that the queries refer to documents in this language and may contain Spanish words or proper names.

The second steps consists simply in the translation of the words that can be found in our English-Spanish dictionary (a fusion of mainly tree dictionaries Vox-Harraps, FreeDict, EuroWordnet). Previously we lemmatize each word using a database of English lemmas.

Each one of the remaining words, that couldn't be translated using the dictionary, are passed as input to the following OOV translation method (based on [2]). Given an OOV English word, we search on Google for Spanish documents at the Web with that word. Our main assumption here is that in a document identified as mainly written in Spanish but that also contains the English word, the window of 20-25 terms around this word (approximately a snippet) might contain its Spanish translation. Taking the first 40 snippets we select the 10 more frequent words (removing stopwords and English words that might also appear). These frequent words are considered as translation hypothesis. To select the final translation we search on Google for English documents containing the transvaluation candidate. If the Spanish word co-occurs in the snippets of this search we count it's frequency and add it to the frequency of the candidate at the first search. This number is used to rank the translation hypothesis, and select the best one.

Finally, the resulting translations are passed as input for the searching process.

4 Searching

In the search phase, we disposed of two ranking options based on the fields identified during the indexation. One possibility was to search over the whole contents of the documents and then rank the results according to this data. The other option was to use the fields identified during the preprocessing. The fields were ordered based on it's descriptiveness of the document [**title, metadata, headings, body, outgoing links**]. In this case, the ranked results are built by the consecutive searches, looking at just one field each time and disposing the results in the same order. At each iteration we remove the duplicated documents at the lower positions. Finally, the

Table 1: Evaluation results

	baseline	**experiment**
Average success at 1	0.0224	0.0821
Average success at 5	0.0672	0.1045
Average success at 10	0.1045	0.1194
Average success at 20	0.1716	0.1343
Average success at 50	0.2612	0.2090
MRR over 134 topics	0.0477	0.0930

process is stopped when the results list arrives to 50 hits of when all the search fields have been explored. In both cases we first try a query with a AND boolean operator for all the words at each query. If the total results are less than 50, we search again with an OR boolean operator.

5 Experiments and results

Our baseline experiment consisted of the query translation as described and searching over the *body* field. Our second experiment follows the same query translation method, and searches over the descriptive fields as explained in the Searching section.

The Mean Reciprocal Rank shows poor results in both experiments, but with a minimum improvement when the descriptive fields are used. Two different factors may be also taken into account. One is the inherent difficulty of the proposed task, and the other is the election of a smaller (but faster) index, based on the language detection program, that possibly has excluded target webpages of the topics.

6 Conclusions and future work

Our main conclusions are:

- The partial indexation of the EuroGov collection seem the first mistake in the design of these experiments. Full indexation should be performed in order to avoid the loss of relevant web pages.
- The strategy of search over the descriptive fields is the best option the known item task, but should also include anchor information to be really effective.
- The translation method for OOV words seems a very promising tool for the Cross-Language Information Retrieval task, but needs more study and a more careful selection of the translation hypothesis.

Our future work will be the improvement of the OOV translation method and its specific testing on different languages. We also will test the descriptive field search strategy adding anchor text from linking webpages. Finally we would like to evaluate the system at the Monolingual and Multilingual tasks of the WebCLEF 2005.

References

[1] Maarten de Rijke Brkur Sigurbjrnsson, Jaap Kamps. Blueprint of a cross-lingual web retrieval collection. *Journal of Digital Information Management*, 3(1):9–13, 2005.

[2] Stephan Vogel Ying Zhang, Fei Huang. Mining translations of oov terms from the web through cross-lingual query expansion. In *SIGIR 2005*, 2005.