Simple Approach to the Spanish-English Bilingual Retrieval Task

Carlos G. Figuerola, José Luis Alonso Berrocal, Angel F. Zazo and Raquel Gómez Díaz

Universidad de Salamanca
Facultad de Documentación
C/ Fco. Vitoria 6-16
37008 SALAMANCA – SPAIN
{figue|berrocal|afzazo|rgomez}@gugu.usal.es

Abstract. This paper describes our participation in bilingual retrieval (formulating queries in Spanish to retrieve documents in English), using an information retrieval (IR) system based on the vector model. Our aim was to carry out a simple approach to solve the problem, without expecting to obtain great results, especially owing to the short time available. The queries formulated in Spanish were translated to English by a commercial machine translation system. The translations were filtered to eliminate stop words, and then the remaining terms were stemmed using a standard stemmer. Results were poorer than those obtained through monolingual retrieval with original English queries, the difference being slightly over 15%.

1 Introduction

This study describes the participation of our team in the Cross-Language Evaluation Forum (CLEF-2000), as first approach to the bilingual information retrieval tasks. Our main objective in participating in the CLEF was to make a first contact in the scope of bilingual information retrieval with Spanish and English, although we have greater experience in monolingual information retrieval in Spanish. Our participation in CLEF 2000 focussed on bilingual retrieval, using queries in Spanish with a collection of documents in English. Obviously, we have also worked with the same queries, formulated originally in English, which have served to establish a line of comparison of results.

The IR problem when more than one language is involved, i.e. evaluating the similarity of a document written in a certain language versus a query in another one, is that of achieving homogeneous representations of both elements (document and query), which may be compared and which may make it possible to establish a degree of similarity between both [6]. Once this homogeneous representation has been achieved, the similarity between a query and each of the documents in the collection can be computed by any of the systems usually used for monolingual retrieval [5]. In our case we use the well-known vector model.

2 Approach to the Problem

For term-based IR techniques, as is the case of the vector model, the terms represented in the documents and in the queries have to be put into the same language. In one way or another, this entails some type of translation, and finding a good translation system can solve the problem.

In principle, it is a matter of translating individual terms, which does not seem to be as complicated as translating a syntactically structured text. However, the main problem, apart from the use of a bilingual dictionary readable by machine, lies in the disambiguation of the terms: these may have diverse meanings and each meaning may have diverse equivalents in the other language. It is not easy to determine the appropriate equivalents in each case and various methods have been proposed for this purpose [1]. The final result depends on the quantity and quality of the semantic knowledge contained in the dictionaries and word lists used.

Thus, we shall not use the approach of translating terms, since this would lead to poorer results in retrieval. Also the translating systems find it easier to disambiguate and contextualize phrases [3], which would give rise to better results.

Hence, and because computationally it is simpler, the process followed was that of translating the queries to the language of the documents, and not the reverse. In our case, a very simple approach was made to solve the problem: that of using one of the commercial machine translation programs. We did not expect great results, although it has allowed us a better understanding of the problem.

2.1 Machine Translation

Although machine translation (MT) is an area of intense research, there are already quite a few commercial programs. These programs do not have much prestige, owing to the fact that the translations obtained often have many mistakes and are sometimes linguistically unacceptable.

But, we noted that the linguistic requirements of vector model based IR systems are not so great as those of the people who have to read and understand translations [4]. Indeed, many IR systems do not examine syntactical constructions and, when the terms are submitted to a stemming process, they disregard morphology.

The use of one of these commercial MT systems does not present any difficulties. In our case, as we lack experience in bilingual retrieval, it seems to be a good way to become introduced to the subject. This was our approach to the problem.

Many MT systems also allow some kind of adaptation to the context, such as domain specific dictionaries, database for language pair translations, etc., which give better results in translation and, consequently, better results in retrieval. However, in our research, none of these additional tools was used. The simplest strategy was followed.

3 The Experiment

The layout of the process followed can be seen in the diagram below:

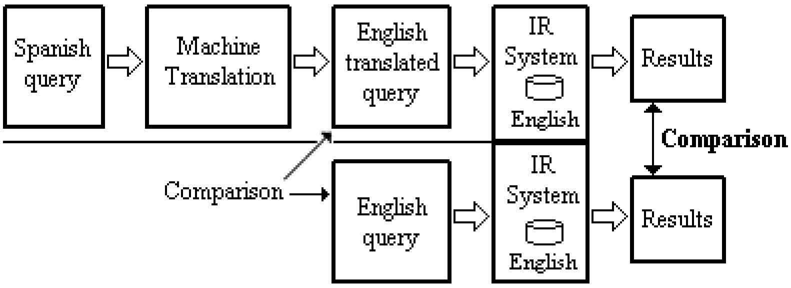

Fig. 1. Spanish-English Bilingual IR system

3.1 Queries in Spanish

We should point out that the queries were not pre-processed, i.e. they were not treated to eliminate terms that might introduce mistakes in subsequent retrieval. Three translation programs were applied directly to the queries in Spanish, without considering the noise that the terms not relevant to the query might introduce into the system.

A future study will be carried out to find out how errors in the translation of the most significant terms in the queries affect information retrieval. We expect to find parallelism between the errors in translation of the queries and the retrieval results.

3.2 Translation of Queries

Three MT programs were used: Systrans (on-line vers. http://www.systransoft.com), Globalink Spanish Assistant v1.0 and Globalink Power translator Pro v6.2. (at present the last two are products of Lernout & Hauspie). These programs are not expensive (the Systrans on-line version is free), and can be used on a PC with few resources.

The reason for using three programmes was to check the quality of the translation, and, consequently, to use the best of the three translations for retrieval. In no case were thematic or contextual dictionaries used. We used the complete topic set in Spanish, i.e. titles, descriptions and narratives, and putted through each of the three translation systems.

The three systems tested produced very similar translations, and also coincided, notably, in the same errors. A study of the errors made by each gave very similar figures for all three. This study was carried out taking into account the significant terms for the retrieval of the original queries in English, contrasted with significant terms of the translations. The different terms were considered as translation errors, except in the cases of evident synonyms. One error was counted in the cases in which

Spanish-English translation produced two or more terms, when in the English queries there was only one. Although this type of count is not very rigorous, it at least allows us to explore the possible differences between the three translation systems tested, always from the point of view of information retrieval.

The error percentages thus estimated were very similar for all three. The differences were very small, with the results obtained by Systran being slightly more favorable. Moreover, and more intuitively, the mere reading of the translations showed that Systran seems to work better with proper nouns. It is better at detecting whether a word is a proper noun, and, when that name can be translated, it also translates it better. Thus, we opted to work with Systran.

3.3 Translated Questions

The translations obtained in the previous phase were processed following the normal retrieval process of the vector model: elimination of stops words, stemming and calculation of weight.

The original queries in English underwent the same treatment. A comparison was made of the stems obtained for the queries translated and those obtained for the original queries in English. A discrepancy of around 28% was observed, i.e. over a quarter of the stems of the queries translated into English were different from the stems of the original questions in English. This does not necessarily mean that the stems obtained were incorrect, since in some cases the translations may have used synonyms, or semantically equivalent terms.

3.4 IR System

As a retrieval engine we used our own software, which we have called Karpanta[1] [2]. This is a simple program based on the vector model, which was designed mainly for educational and not operational purposes. Owing to the large number of documents used in the experiment (113,000 documents, 400 MB of information) the operation process was frustratingly slow. This did not worry us at first, since the objective of our study was to verify the use of a simple approach to the problem: the application of an inexpensive MT system to CLIR.

Before indexing the documents in English, stop words were eliminated in order to save index space. For this purpose a standard list of some 200 components was used. Remaining words were stemmed by applying Porter's algorithm [PORTER80]. We used a Perl script with an implementation of this algorithm, which is widely diffused through CPAN [7]. Karpanta was then used to index all the documents in English, with all their fields. The weights of the stems obtained were calculated with the usual scheme of frequency of term in the document by *IDF*.

The queries translated into English were processed in the same way. They were used as a whole, with title, description and narrative; stop words were eliminated and

[1] A legendary figure in Spanish comics, whose most outstanding characteristic was that of always being hungry.

stems obtained whose weight was calculated in the same way. The solving of the queries, i.e. the computation of similarity between each query and each of the documents, was performed with the widely known and used cosine formula.

The same process was also followed for the original queries in English, thus obtaining results, which have served as a reference point to establish comparisons with the results obtained after bilingual retrieval.

It should be emphasized that in no case was a relevance feedback used in our experiments, despite the fact that this would probably have given rise too much better results.

4 Results

The results obtained with the queries translated from Spanish gave a mean precision of 0.2273 and can be seen in the attached graph. However, the results were fairly unequal from some queries to others (standard deviation = 0.23).

If we compare these results with those obtained using the original queries in English (mean precision of 0.27), the former are slightly lower. The precision-recall curves are almost parallel.

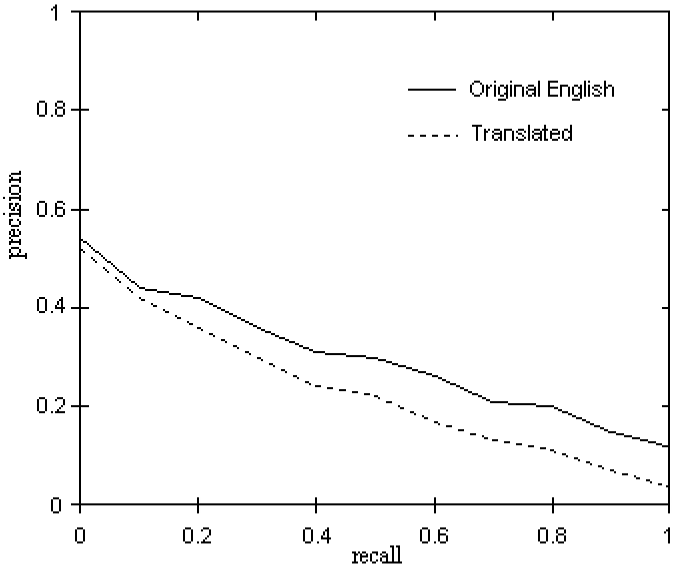

Fig. 2. Spanish-English Bilingual Retrieval Comparison

Moreover, if we observe each individual query, it can be seen that there are many parallels: the queries translated into English which give the best results coincide with the original queries in English that work best. Those with the worst results are also the same, both in the original queries in English and in those translated into Spanish.

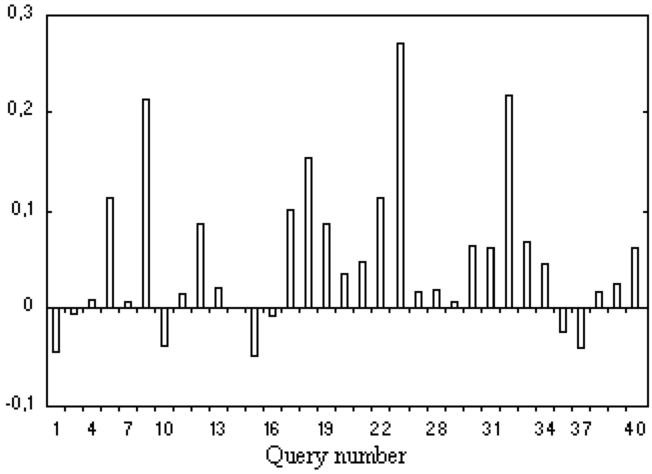

Fig. 3. Difference in mean average precision from the English original queries set and translated one.

5 Conclusions

The use of a commercial MT system to solve bilingual retrieval tasks is an easy and swift solution, although effectiveness in retrieval is slightly below that obtained in monolingual results. The difference is around 15%, although this figure is less at low recall levels, i.e. taking into consideration only the first documents retrieved.

No relevance feedback of queries was performed in our experiments, although this would probably have led too much better results.

Future work will be done to find out how translation errors of significant terms for the retrieval affect the results.

References

1. Agirre, E., Atserias, J., Padró, L. and Rigau, G.: Combining Supervised and Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation Computers and the Humanities. Special Double Issue on SensEval. Eds. Martha Palmer and Adam Kilgarriff. 34:1,2, (2000). [http://www.lsi.upc.es/~nlp/papers/chum99-arpa.ps.gz]
2. Figuerola, C.G., Alonso Berrocal, J.L. and Zazo, A.F.: Diseño de un motor de recuperación para uso experimental y educativo. BiD: textos universitaris de biblioteconomia i documentació, 4 (2000) [http://http://www.ub.es/biblio/bid/04figue2.htm]
3. Fluhr, C.: Multilingual Information Retrieval. In Cole, R. A. et al.: Survey of the Sate of the Art in Human Language Technology, Standford University, Stanford, CA,(1995) 391-305 [http://www.cse.ogi.edu/CSLU/HLTsurvey/ch8node7.html]

4. Hull, D.A. and Grefenstette, G.: Queryng Across Languages: A Dictionary-Based Approach to Multilingual Intormation Retrieval. SIGIR96 (1996) 49-57

5. Kowalski, G.: Information Retrieval Systems - Theory and implementation. Kluwer Academic Publishers (1997)

6. Oard, D. and Dorr, B.J.: A Survey of Multilingual Text Retrieval. (1996) [http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps]

7. Phillips, I.: Porter's stemming algorithm. Perl script. http://www.perl.com/CPAN-local/authors/Ian_Phillipps/Stem-0.1.tar.gz

8. Porter, M.F.: An algorithm for suffix stripping. Program, 14(3) (1980) 130-137

9. Systran Software: SYSTRAN-Translation Technologies, Language Translator, Online dictionary, Translate English (2000) [http://www.systransoft.com]