Retrieval of bilingual Spanish-English information by means of a standard automatic translation system

Carlos G. Figuerola *(figue@gugu.usal.es)*
José Luis Alonso Berrocal *(berrocal@gugu.usal.es)*
Angel Francisco Zazo Rodríguez *(afzazo@gugu.usal.es)*
Raquel Gómez Díaz *(rgomez@gugu.usal.es)*

Universidad de Salamanca
Facultad de Documentación
c/ Fco. De Vitoria 6-16
37008 SALAMANCA-ESPAÑA

ABSTRACT:

This paper describes our participation in bilingual retrieval (queries in Spanish on documents in English), by means of an information retrieval system based on the vector model. The queries, formulated in Spanish, were translated into English by means of a commercial automatic translation system; the terms extracted from the resulting translations were filtered in order to get rid of empty words and then they were normalised by stemming. Results are poorer than those obtained through monolingual retrieval with the original queries in English slightly above 15%.

Introduction

Our participation in CLEF 2000 is centred mainly on bilingual retrieval, by which we mean queries in Spanish against a collection of documents in English. Obviously, we have worked with the very same queries formulated originally in English, which we have used to obtain a line of comparable results.

When one tries to solve queries in a given language versus documents written in a different one, the problems is to get a homogeneous representation of both queries and documents, so that they can be compared and thus allowing us to establish a measure of similarity between them [OARD96]. Once this homogeneous representation has been obtained, the similarity between a query and each of the documents of the collection may be computed by means of any of the systems usually employed for monolingual retrieval.

For term-based information-retrieval systems, as is the case of the vector model [SALTON83], the question is to insure that the terms that represent documents and queries use the same language. One way or another, this involves some sort of translation; at least in principle, translation of queries seems to be less expensive than translating whole documents. Anyway, the problem is the translation of individual terms, which seems less complex than translating a sintactically structured text. The main problem,

beyond using a machine-readable bilingual dictionary, is to disambiguate those terms: each of them may have different meanings and each of them will have a different equivalent in the other language. It is not easy to determine the proper equivalents for each case and several methods have been proposed with this purpose [AGIRRE2000]; final results depend, to a large extent, of the quantity and quality of semantic knowledge contained in the lexicons and dictionaries employed .

An obvious alternative to approach the problem of bilingual retrieval is to use some automatic translation system; there is quite a number of commercially available systems. However, these systems are not too well liked, since in general terms the translations they produce contain many mistakes and, occasionally, are not acceptable from a linguistic point of view. It must be noticed, however, that the linguistic requirements of retrieval systems are rather lower than those of the persons who must read and understand the translations [HULL96]. In fact, many information-retrieval systems do not use or consider syntactic constructs and, when terms experience some kind of normalisation process, they ignore morphology.

The utilisation of one of these automatic translation commercial systems poses no difficulties and, in our case, lacking experience in bilingual retrieval, seems to be a good way to start on this subject.

Experiment

The retrieval engine we have is used is our own software, which we call Karpanta [FIGUEROLA2000]. It is a simple program, based on the vector model, and it has been designed with educational (vs. productivity) purposes. It works, although it is rather slow for large numbers of documents. On the other hand, the goal of our work is to check the efficiency of a standard automatic translation system when it is applied to information retrieval; rather than as a monolingual retrieval technique.

Hence, we used Karpanta to index the whole lot of documents (in English), keeping all of their fields. We had eliminated empty words previously, using a standard list of empty English word that consists of approximately 200 words.

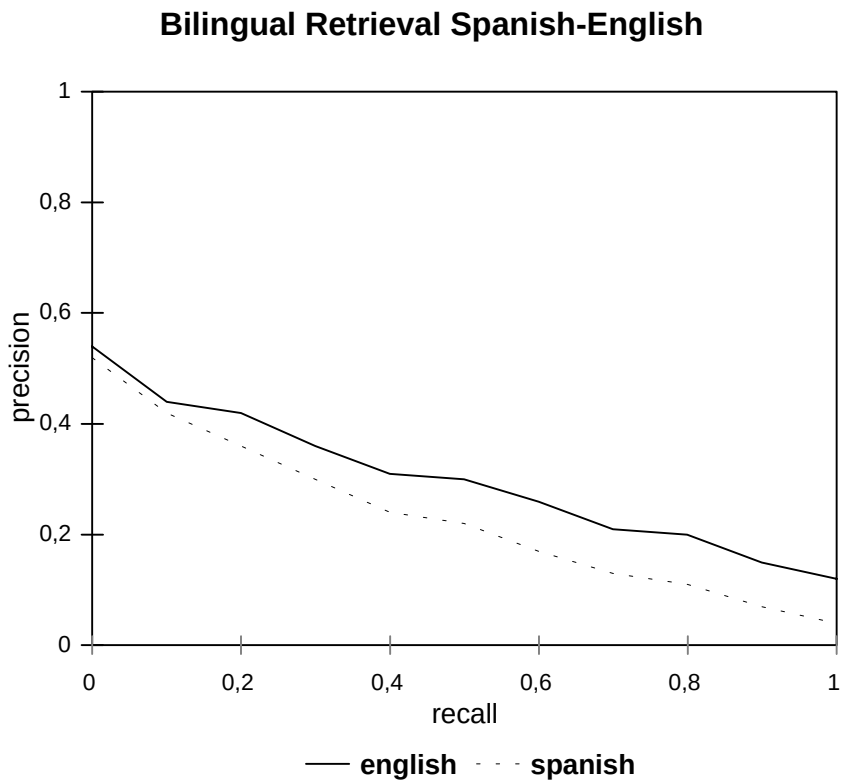
Non-empty words were stemmed by means of Porter's well-known algorithm [PORTER80]. This was done by means of a Perl script that implements the above algorithm; this script has been spread widely by CPAN [PHILLIPS95]. The weights of the terms or stems we obtained were calculated by means of the usual scheme of term frequency in the document x IDF.

The original queries, in English, were dealt with in the same way. We used all of their fields, empty words were eliminated and thus we obtained the stems, whose weight we measured as before. Query resolution, that is, the computation of similitude between each query and each document, was made by means of the usual cosine formula; thus we obtained the results we have used as the basis to establish comparisons with the results we got afterwards in bilingual retrieval.

Spanish queries were translated into English by means of an automatic translation system. Actually, we tried various commercial systems: Systran, Power Translation Pro, Spanish Assistant. Although most automatic translation systems allow for some kind of context adequacy and training, using such things as specific lexicons, translation memories etc., we did not use any of these possibilities. In the case of Systran, we actually used the web-accessible version [SYSTRAN2000]. All the systems we tried produced rather similar translations; they also tended to produce remarkably similar errors. We finally gave the nod to Systrans, since it seems to have better capabilities when recognising proper nouns; besides, it is better at translating them, when at all possible.

The translations thus obtained were processed in the same way as the original queries in English: elimination of empty words, stemming, computation of weights and calculation of similitude with each document.

A comparison between the stems produced for each of these translations and those produced by the original queries in English shows the divergences. If we compare a list of the stems we obtained by means of the original queries in English with those obtained in translations, we observe that an average of 28% are different. This does not mean they are necessarily incorrect since in some cases the translations may have used synonymous or semantically equivalent terms.



Results

The results we have obtained with queries translated into Spanish produce an average accuracy of 0.2273 and they have been shown in the previous graph. However, results show rather large variations between queries (typical deviation=0.23).

On the other hand, if we compare these results with those obtained from the original queries in English (with an average precision of 0.27), they are clearly inferior. Precision-Recall curves are almost parallel. However, if we examine each individual query, it can be seen that the ones that produce the best results in English are also the ones that work best in the Spanish-to-English translation. Similarly, the queries that produce the worst results are also the same, both in the original (English) queries as in the queries translated into Spanish.

Conclusions

The use of a standard system of automatic translation to solve bilingual retrieval tasks is an easy and fast solution, although the efficiency we achieved in retrieval is clearly lower than the one obtained by means of monolingual queries. This reduction is about 15%, although it is lower for reduced levels of completeness (that is, taking into account just the first few documents we find).

REFERENCES

[AGIRRE2000] 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]

[FIGUEROLA2000] Figuerola, C.G.; Alonso Berrocal, J.L. & Zazo Rodríguez, A.F.: ”Diseño de un motor de recuperación para uso experimental y educativo”, BiD: textos universitaris de biblioteconomia i documentació, 4 [http://http://www.ub.es/biblio/bid/04figue2.htm]

[HULL96] Hull, D.A. & Grefenstette, G.: " Queryng Across Languages: A Dictionary-Based Approach to Multilingual Intormation Retrieval", SIGIR 96, 49-57

[OARD96] Oard, D. & Dorr, B.J. : "A Survey of Multilingual Text Retrieval",[http://www.clis.umd.edu/dlrg/filter/papers/mlir.ps]

[PHILLIPS95] Phillips, I. http://www.perl.com/CPAN-local/authors/Ian_Phillipps/Stem-0.1.tar.gz

[PORTER80] Porter, M.F.: “An algorithm for suffix stripping”, Program, 14(3), 130-137

[SALTON83] Salton, G. & McGill, M. : Introduction to Modern Information Retrieval, New York, McGraw-Hill, 1983

[SYSTRAN2000] Systran Software: SYSTRAN - Translation Technologies, Language Translator, Online dictionary, Translate English, [http://www.systransoft.com]