IR-n, a passage retrieval system from University of Alicante, at Clef 2001.

Fernando Llopis y Jose Luis Vicedo
{llopis,vicedo}@dlsi.ua.es
Depto. Lenguajes y Sistemas Informáticos. Universidad de Alicante
Campus de San Vicente del Raspeig
Apartado 99, 03080 Alicante, Spainj

Abstract
Previous works showed that the use of document passages like basic unit of information, to calculate the relevance of a document to a question, improve the results of the information retrieval systems sensibly. However, IR community has not arrived to a consent about how to define those text passages so that the system can improve the efficiently. This paper reports on experiments with **IR-n** system, a information retrieval system based on the selection of passages of variable size as basic unit of information, in the monolingual (Spanish) and bilingual (Spanish-English) tasks at CLEF-2001. The IR-n system has been developed this year in the Language Processing and Information Systems research group at the University of Alicante.

1 Introduction

Information retrieval systems (RI) has as the main objective select, from a collection of documents, most relevant ones for a certain question. These systems measures the level of similarity between a document and the question. The frequency of appearance of the terms of the question in the document is very important to calculate this similarity. This can cause, in texts of considerable size mainly, that appearance high frequencies of some terms of the question, propitiates that a document can be considered relevant without being really.

An alternative to this model is the proposal of systems of RI that values the relevance of the documents in function of the relevance of the fragments or passages that forms them, where each passage is a group of contiguous text inside the document. This approach denominated passage retrieval, (PR), allows that the calculation of similarity is not affected excessively by the size of the document and can determine with more precision the part of the document that is more relevant to the question (It's very important when the document has a great size).

Systems that uses technical of PR are more complex than systems of traditional RI. First, because they require store a more quantity of information for each one of the terms in the document (usually the position that occupies in the document) and in second place, that the number necessary calculations to evaluate the relevance of each one of the passages of the document is higher. Nevertheless, evaluations carried out in other works [1][2][8] reflects that the increment of complexity is rewarded with better results.

In [2] the types of passages are divided in three classes: discourse (based upon textual discourse units); there is a type classification, Semantic (based upon the subject or content of the text) or window (based on the number of words).

The IR-n system that develops this work is included in the models of PR based on the discourse The system uses passages of variable size that are defined based on a certain number of sentences. The passages are generated in overlapping way in the document, that is to say, if the size of the paragraph is N, the first paragraph will be formed from sentence 1 to N, the second from 2 to N+1 and so on. This way, the similarity of each one of the passages of a document with the question will be evaluated and finally this document will be punctuated with the best valuation that any of the passages that forms it has obtained.

This paper is structured in the next way, in the following section IR-n system architecture is described. Afterwards we analyse the results obtained by a the different test we do at clef 2001. Finally we extracted initial conclusions and open directions for future work .

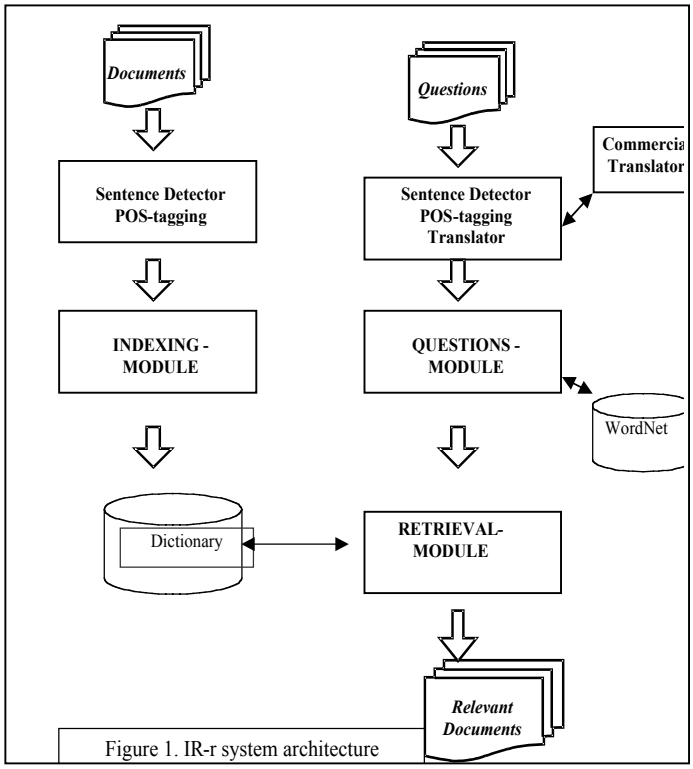


Figure 1. IR-r system architecture

2 System overview

IR-n system has been implemented in C++ in a cheap Linux box. We have good times in process of indexing and retrieval.

IR-n system is structured in three modules: Indexing module, Questions module and Retrieval module. First module processes and indexes all the collection of documents; Question module processes the question and expands or translates, if it is necessary. Retrieval module ranks the documents according to a similarity of a documents and query. Figure 1 shows the architecture.

2.1 Indexing module

This module has as main objective the generation of the dictionaries with the necessary information to use in the retrieval process. The indexing terms consists of character strings made up of letters, numbers and symbols whose length is minor than 21 characters. Previous to their indexation documents are pre-processed to detect sentence boundaries, and part of speech tagging terms. Most frequent terms are eliminated using of a list of words (stop-words) also .

The system requires storing for each term, the number of documents where each term appears and for each one of the texts the number of appearances of each term in the text and the position of each word in the text (sentence number and order inside the sentence). This supposes an increment of the information to store regarding those systems of information retrieval based on complete documents.

For each term we store the stem when we work with English documents and lemma in Spanish documents. It is due to we think that in Spanish the stem may not be relevant.

2.2 Question module

In bilingual task, the first step of question pre-processing consists on translating the topic from Spanish to English using a commercial translator. We want to test if it is possible to use commercial translators and to have good results.

The following step eliminates stop-words ant detects stem (English topics) or lemma(Spanish topics) in both tasks.

Query expansion using semantic information was used in one of the tests. Basically consists in obtaining the synonyms for each term of topic, using Wordnet, a lexical thesaurus.

2.3 Retrieval module

This module is the one in charge of recovering the documents in function of its similarity with the question. The process of measure the similarity of each document and presentation of results is the following one:

1. Order the question terms from smaller to larger in function of the number of documents of the collection in those that they appears.
2. Obtain the documents that contains at least one term.
3. Calculate the value of similarity of each document
4. Order the documents in function of their similarity to the question.
5. Visualization of the results in form of orderly list.

To calculate the similarity of topic with a document, system calculate the similarity of topic with each passage of document (where a passage is a number of sentences contiguous in the document) first and after system assigns the document the highest value in similarity of the passages that forms it

The similarity of topic with a passage is calculated in the next way.

Similarity of the passage = $\sum_{t \in p \wedge d} W_{p,t} * W_{q,t}$

Where:

Wp,t = log(fp,t + 1). being fp,t the number of appearances of term t in passage p.

Wq,t = loge(fq,t + 1) * idf. Being fq,t the number of appearances of term t in question q.

idf = loge(N / ft + 1). being N the number of documents of the collection, and ft is the number of different documents where the term appears t.

As it can be observed, the formulation used to value the similarity between each passage and the question is similar to the measure of the cosine [10]. The only difference is that the normalization that uses this measure is omitted, we think that this normalization is not necessary due to the size of passages is not so much different between them.

One of the main things for us was to determine the number of sentences to improve the results. For it, we test passages of several sizes (in number of sentences) in the collection and topics of past year. The results can be seen at table 1.

	Precision in Passage retrieval					
Recall	5 sentences	10 sentences	15 sentences	20 sentences	25 sentences	30 sentences
0.00	0,6378	0,6508	0,6950	0,7343	0,6759	0,6823
0.10	0,5253	0,5490	0,5441	0,5516	0,5287	0,5269
0.20	0,4204	0,4583	0,4696	0,4891	0,4566	0,4431
0.30	0,3372	0,3694	0,3848	0,3964	0,3522	0,3591
0.40	0,2751	0,3017	0,2992	0,2970	0,2766	0,2827
0.50	0,2564	0,2837	0,2678	0,2633	0,2466	0,2515
0.60	0,1836	0,1934	0,1809	0,1880	0,1949	0,1882
0.70	0,1496	0,1597	0,1517	0,1498	0,1517	0,1517
0.80	0,1213	0,1201	0,1218	0,1254	0,1229	0,1279
0.90	0,0844	0,0878	0,0909	0,0880	0,0874	0,0904
1.00	0,0728	0,0722	0,0785	0,0755	0,0721	0,0711

Table 1

It can be observe that better results are obtained when the passages are formed by 20 sentences, and then this is the size we choose for Clef 2001 experiments.

We take another measure for reducing the memory requirements and execution time. In the step of the obtain the documents that contains at leas one term of the topic, we work with a limitation, only adding new documents until arrive to 5% of the number of total documents. This measure is mentioned as efficient in [1], we have also corroborated that increasing these percentage sensitive improvements is not obtained in the results.

3. Experiments and results

This year we have participated in two tasks in clef 2001, bilingual (Spanish-English) and monolingual (Spanish). The bilingual task consists in querying in Spanish a document collection of English texts. Monolingual task consists query a Spanish collection in Spanish. The test collection for CLEF 2001 consists of SGML-formatted documents from national newspapers (Los Angeles Times – 1994) for the first task and news agency (Agencia EFE S.A. Spanish news agency- 1994) for the second task, main topic set consists of 50 topics and is prepared in Spanish.

We have carried out 3 runs in monolingual task (called EI, PR, PRM) and 4 runs in bilingual task(called EI, PR, PRM ,and EXP).

3.1 Runs descriptions

EI run. This run calculates the similarity of the topic with each document using a standard method of information retrieval (cosine measure [10]) based on complete documents,

PR run. This run uses the method proposed in IR-n system.

EXP run. This run was used only for bilingual task. It consists in adding a topic, the main synonyms for each term and after execute the retrieval process proposed in IR-n system.

PRM run. In this run we determine the sentence boundaries in topic (including title, description and narrative)first , we calculate for each sentence of topic the measures of similarity (like PR run) and we order the relevant documents in function of the mean of measures of similarity of each sentence. This measure gives more importance to title, due to terms of title usually appears in narrative too.

The first run uses title and description of the topic, second and third only use only title and the fourth use all terms from the topic.

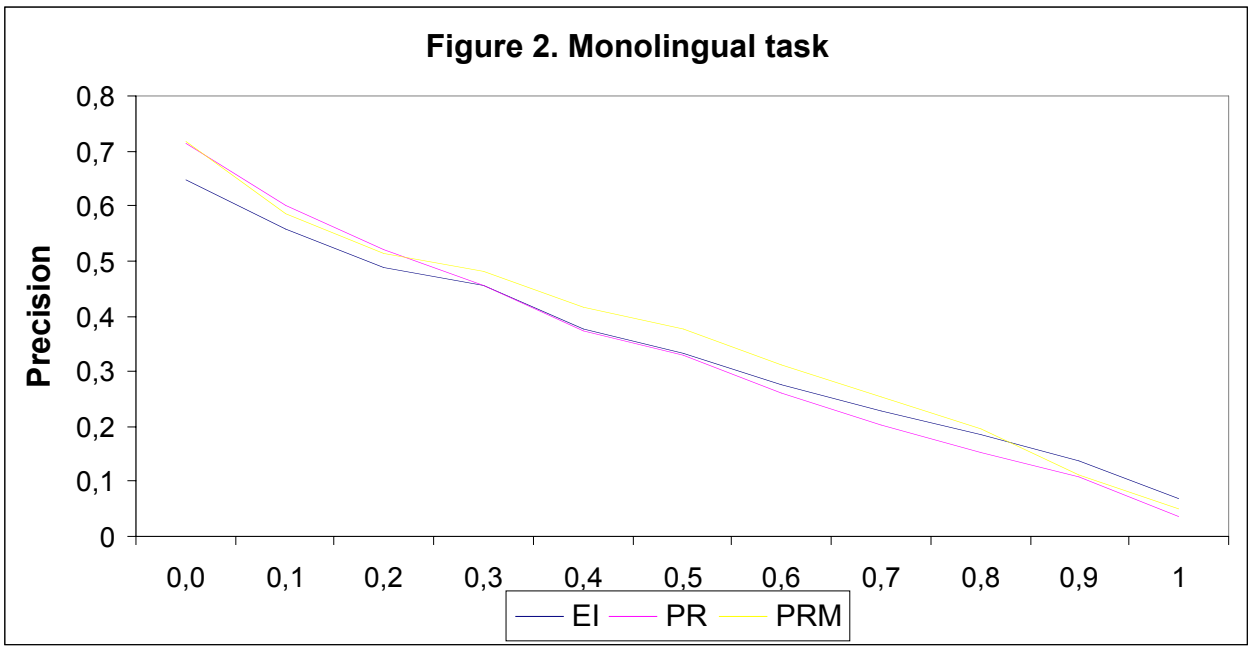
3.2 Results in Monolingual Task

In monolingual we have carried out three tests (EI; PR and PRM).

In this task, the results using passage retrieval (PR and PRM runs) give no significant improvement than using Information retrieval with full documents. We think that this results may be due to the size of most of documents of the collection.

The results using short query (PR run) are comparable using long query (PRM run). It may be due to then most of the relevant documents that includes narrative terms, including title terms too..

Figure 2 depicts the results reach with these tree runs

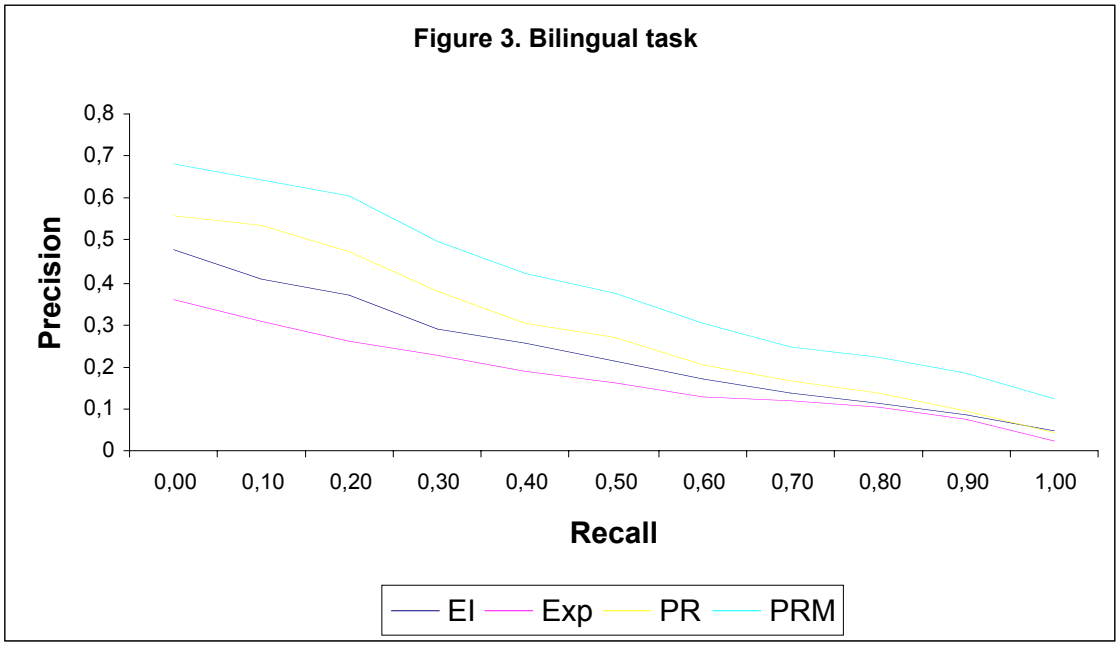


3.3 Results in Bilingual task

In monolingual we carried out four tests (EXP, EI, PR and PRM).

We have obtained the better results with the PRM test. In bilingual task we have a significant improvement using passage retrieval. The sensible difference, in this case, between using long or short queries (PRM and PR runs) may be due to many relevant documents containing the terms of topic narrative and no the terms of title. Another important aspect is the bad result we have obtained using query expansion with semantic information. The fact of expand the topic terms has produced that the result has even been worse than the standard measure.

Figure 3 depict the results reached with these four tests



The comparative of the results of several runs in both tasks has demonstrate us that the line we choose (Passage retrieval) is better than work with full documents.

3.4. Results of IR-n system al clef 2001

Tables 2 and 3 depicts the results of our runs and the median results obtained by the other systems at clef 2001. In these tables we report the median average precision for all the individual queries.

Monolingual task	Average precision	Increment
Median	0.4976	0.0
Runs		
PRM	0,3528	-28.09
PR	0,3287	-33.94
EI	0,3297	-33.74

Table2 . IR-n at clef 2001 Monolingual task

In table 2 we have realized that our results are worse than the median in clef 2001. We think that our result may be due to the incorrect election of a tagger we use for Spanish we use. In experimental test we detect many errors in the process of lemma election.

Monolingual task	Average precision	Increment
Median	0.2422	0.0
Runs		
PRM	0,3759	+55.20
PR	0,2725	+12.514
EXP	0.1672	-30.96
EI	0,2197	-9.28

Table3 . IR-n at clef 2001 Bilingual task

In this bilingual task IR-n system has obtained a sensibly better result than the median at clef 2001. We have realized that the behaviour of system in monolingual and bilingual are really similar. Also, it is possible that if we don't have so many errors in the process of tag the Spanish documents , the result in monolingual task had been better than bilingual task.

Nevertheless it is strange the median at clef 2001 is very different between both tasks.

5 Conclusions and future work

As the results demonstrate , use of groups of sentences like basic text unit for the measure of the similarity between questions and documents in the environment of a RI systems, has been revealed as a very effective technique. Nevertheless it is possible than the selection of the appropriate size of passage depends on the type of documents collection. In bilingual task (La Times) we have obtained a considerable improvement when we use Passage Retrieval techniques, but this improvement is small when we work with the document collection in monolingual task (Efe).

Also, the type of collection is significant in respect of the type of question. The results reach using title+narrative are better than using only title + description, however the improve reached is no significant in Spanish collection and is significant in English collection.

The use of techniques of query expansion has reached sensible worst results than the rest of test, It is possible that using this techniques with passage retrieval directly, giving the same weight at all the terms (base and expanded terms) it is no effective.

The use of a commercial translator in the bilingual task, without manual supervision, does not seems being a bad election. In fact our results in Bilingual task are over the median results.

After this first experience we open several lines of future work. We want to study the number of sentences that should conform a paragraph to improve the results, and possibly determine this number in function of the type of question or document collection. We think that optimise the searching process with the intention of reducing the temporary complexity of the process is important .In spite of the bad results reached using query expansion we want to continue studying the advantages and inconveniences of carrying out a process of expansion of the questions.

References

[1] M Kaszkiel, J Zobel and R. Sacks-Davis. Efficient Passage Ranking for Document Databases. ACM transactions on Information Systems, Vol 17, Nº 4, October 1999, Pages 406-439
[2] J.P. Callan. Passage-level evidence in document retrieval. In Proceedings of the 17 th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland, 1994, pp. 302-310.
[3]F. Crivellari, M. Melucci. Web Document Retrieval using Passage Retrieval, Connectivity Information, and Automatic Link. In proceedings of the ninth Text REtrieval Conference(TREC-9).
[4]I. Namba Fujitsu Laboratories TREC9 Report. In proceedings of the ninth Text REtrieval Conference(TREC-9).
[5] Hearst, M. and Plaunt, C. (1993). Subtopic structuring for full-length document access. In Khorfage, R., Rasmussen, E., and Willet, P., editors, Proceedings of the 16th ACM-SIGIR conference, pages 59--68, Pittsburgh, USA.
[6] Hearst, M. 1994. Multi-Paragraph Segmentation of Expository Text. Procs. 32nd Annual Meeting of the Assoc. for Computational Linguistics (ACL-94), pp. 9-16.
[7]I.H. Witten, A. Moffat, T. Bell Managing Gygabytes. Ed morgan Kaufman 1999
[8] M. Kaszkiel, J. Zobel Passage Retrieval Revisited SIGIR '97: Proceedings of the 20th Annual International ACM July 27-31, 1997, Philadelphia, PA, USA
[9] J. Callan, B. Croft, and J. Broglio, "TREC and TIPSTER Experiments with INQUERY," Information Processing & Management, 31(3):327-343, 1994..
[10]Salton G.(1989), Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison Wesley Publishing, New York.