Cross-language experiments with IR-n system*

Fernando Llopis and Rafael Muñoz
Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información.
Departamento de Lenguajes y Sistemas Informáticos.
University of Alicante, Spain
{*llopis,rafael*}*@dlsi.ua.es*

Abstract

This paper describes the third participation of IR-n system at CLEF-2003. Two previous participation are focused on Spanish monolingual task. This year, we participated in three different tasks: multilingual task (four languages), bilingual task (Italian-Spanish) and monolingual task (Spanish, German, French, Italian). This paper describes the experiments carried out as training process in order to set up the main system features and shows the results obtained. These results shows that IR-n system obtains good scores for the three tasks improving the average of the CLEF 2003 systems.

1 Introduction

Information Retrieval (IR) systems have to find the relevant documents to a user's query from a document collection. We can find different kind of IR system at the literature. On the one hand, if the document collection and the user's question are written in the same language then the IR system is a monolingual system. On the other hand, if the document collection and the user's question are written in different languages then the IR system is a bilingual (two different languages) or multilingual (more than two languages) system. Obviously, the document collection for multilingual system is written in two different languages at least.

This paper presents the adaptation of IR-n system [7] to participate at CLEF'2003. Our system participates in the following tasks:

- monolingual tasks:
  - Spanish
  - French
  - German
  - Italian
- bilingual tasks:
  - Italian-Spanish
- multilingual tasks:
  - Spanish-German-French-Italian

*This work has been partially supported by the Spanish Government (CICYT) with grant TIC2000-0664-C02-02 and (PROFIT) with grant FIT-150500-2002-416.

IR-n system is an IR system based on passages instead of traditional systems based on full document. Every passage is made up of a fragment or piece of text [1, 6]. Theses systems calculates the document's relevance studying their passage's relevances. IR-n system calculates the similitude between user's query and documents using a set of passages.

This proposal adds the following advantages:

- To consider the proximity of appearance of query terms into the documents.
- To define a new information transmission unit, more adequate both users than further treatment.
- To avoid normalization problems of documents.

2 IR-n: a information retrieval system based on passages

This section presents the conceptual modelling of IR-n. The following main features are presented in the next subsections:

1. Passage concept.
2. Similarity measure between the user's question and the documents collection
3. Similarity measure between the user's question and the documents collection based on similarity passages
4. The use of query expansion in the IR-n system

2.1 Passage concept

First Passage Retrieval systems (PR) used the paragraph as passage size. The use of paragraph as passage unit caused the built of an heterogeneous collection of passages due to the different size of paragraphs. Moreover, this segmentation do not guarantee the fact that every passage is related to different subjects. For this reason, further proposals of PR systems used more than one paragraph as passage.

A different trend proposes the use of a number of words as passage [6, 1]. These proposals solve the heterogenous size problem of previous PR systems. Moreover, this kind of system can to adapte easily the number of words to the document collection and the user's query. This flexibility is very important to increase the performance of systems [6]. However, these systems lose the syntactic structure. In fact, a passage made up of 100 words can including one or two incomplete sentences.

Between the word and the paragraph exists an unit with structure that it is the sentence. The main aspect of the sentence is the self-contained of meaning. This aspect is very important in Information Retrieval because the answer of system can be understood by the user. Obviously, an only sentence does not have an enough identity to determine if a document that contains it is relevant in relation to certain topic. Although it establishes some limits and helps to value the fact that terms of the user's query appear in the same sentence. Since the sentence does not have an entity the sufficiently complete to define a passage, the passages are defined as a set of consecutive sentences. The system IR-n uses the sentence as basic information unit to define the passages. The size of passage can be adapted to improve the efficiency of the IR-n system. The size of passage is measured in number of sentences. The use of sentence to define passages presents advantages against the use of paragraph or word.

The use of paragraph as unit to define the passage has two main problems:

- It is possible that the documents collection does not have information about the paragraphs marks in the document.

- The paragraphs can be used for visual reasons instead of structural reasons of the document.

The use of a number of words as unit to define the passage presents two problems:

- The number of words to be considered as a passage depends on the writing style used. The same event is described using less words in a document of news agency that in a newspaper. If the same event is also written in a novel the number of words will be bigger.
- If the system uses words to define the passage, it can happen that the lack of structure of the considered text fragment can cause the does not understanding of the text recovered. This is due to the fact that the passage can start and end in any part of document.

Finally, the use of sentences to define the passage presents the following advantages:

- A sentence usually expresses an idea in the document.
- Usually documents use the punctuation signs to separate ideas. There are algorithms to obtain each sentence from a document using their superficial structure with a precision of 100% [10].
- Sentences are full units allowing both to show an understandable information by users, or to provide this information to a subsequent system (for example a system of Question Answering). For this reason, the use of sentences improves the proposals that define the passages using a number of word.
- The use of sentences to define the passage allows to work with a heterogeneous document collection written by different authors an with different literary styles. Moreover, the size is a parameter of the system easily adaptable to the language, kind of texts, the size of the user's query, or the final use of the recovered passages. In this case, it is similar to the window model that can re-size the wide of window depending on the document collection.

2.2 Similarity measure between passage and user's question

At the beginning, the system IR-n used the traditional measure of the cosine [11]. Nevertheless, further experiments carried out using other similarity measures obtained better results. The similarity measures of the system IR-n differs from traditional IR system that IR-n system does not use the normalization factors related to the passage or document size. This is due to the fact that passage size is the same for all documents. So, the system IR-n calculates the similarity between a passage P and the user's query q in the following way:

$$sim(Q, P) = \sum_{t \in Q \wedge P} (w_{Q,t} \cdot w_{P,t}) \tag{1}$$

where:

$$w_{Q,t} = freq_{q,t} \cdot \log_e(\frac{N - freq_t}{freq_t}) \tag{2}$$

$$w_{P,t} = 1 + \log_e(1 + \log_e(freq_{p,t} + 1)) \tag{3}$$

where $freq_{Y,t}$ is the number of appearances or the frequency of the term t in the passage or the question Y. N is the total number of documents in the collection and $freq_t$ is the number of different documents that contain the term t.

2.3 Similarity measure of document based on similarity passages

All systems of PR calculate the similarity measure of the document in function of the similarity measure of their passages using the sum of similarity measure for each passage or using the best similarity measure of passage for each document. The experiments carried out in [5] have been ran by the IR-n system, obtaining better results when using the best similarity measure of passage like similarity measure of the document.

Our proposal is based on the fact that if a passage is relevant then the document is also relevant. In fact, if a PR system uses the sum of every similarity measure of passage the the system has the same behavior like IR system based on document adding concepts of proximity.

Moreover, the use of the best similarity measure of passage allows to obtain the best passage improving further search process.

The system IR-n calculates the similarity measure of the document based on the best similarity measure of their passages in the following way:

$$sim(Q, D) = \max_{\forall_{i:P_i \in D}} sim(Q, P_i) \tag{4}$$

2.4 Query expansion

The techniques of query expansion allow to locate relevant documents that do not contain the exactly words of the user's query.

Different studies have been carried out in order to add these techniques to the system IR-n. These studies were two: the former, in the CLEF-2001 [9], the system added the synonyms of the terms of user's question. This experiment achieved lower results than the IR-n system without question expansion. The latter, in the CLEF-2002 [8], a model of Relevance Feedback was proposed achieving a few better results.

This year, the pattern proposed in [3] has been used, but adapting to the passage retrieval. This algorithm increases the relevance of each added term because are closely to the remaining term of question in the document.

3 Training process

This section describes the experiments carried out in order to obtain and to optimize some system's features to improve the performance of the system. The training corpus used in these experiments was the Clef-2002 document collection. Moreover, all experiments have only been carried out using short questions, that is the system only used the title and description from the query. The following subsections explain the specific experiments carried out to every CLEF task.

3.1 Monolingual experiments

First experiments are focused on establishing the adequate number of sentence (N) to make up the passage for each language (Spanish, Italian, German, French and English). The performance of the system was measured using the standard average interpolated precision (AvgP). For every language, the stemmers and the stop-word lists used were provided by the organization of the clef[1].

Table 1 shows the scores achieved for each language without query expansion. The German scores are also obtained without splitting the compound nouns. The obtained figures show the best results for German, French and English using 14 sentences, for Spanish using 9 sentences and for Italian using 8 sentences. The bigger size for German, English and French is due to the kind of document collections used for each language. The three collections are made up of documents with a bigger number of sentences that Spanish and Italian documents. Moreover, the lowest scores achieved for German language (0,3364) show the influence of splitting the compound nouns. The

[1] http://www.unine.ch/info/clef

	Passage size using number of sentences										
Size	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Spanish	0,4839	0,4974	0,5015	0,5004	**0,5042**	0,5001	0,4982	0.4978	0.4973	0.4973	0.4983
Italian	0,4073	0,4165	0,4171	**0,4207**	0,4146	0,4190	0,4188	0,4193	0,4195	0,4166	0,4158
German	0,3236	0,3278	0,3268	0,3267	0,3287	0,3293	0,3315	0,3327	0,3350	**0,3364**	0,3363
French	0,4260	0,4347	0,4442	0,4519	0,4529	0,4625	0,4655	0,4685	0,4716	**0,4731**	0,4725
English	0,4675	0,4697	0,4800	0,4883	0,4882	0,4957	0,4923	0,4945	0,4979	**0,5057**	0,5038

Table 1: AvgP without query expansion

	Passage size using number of sentences										
Size	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
No Splitted	0,3236	0,3278	0,3268	0,3267	0,3287	0,3293	0,3315	0,3327	0,3350	**0,3364**	0,3363
Splitted	0,3843	0,3894	0,3936	0,3933	0,3972	0,3982	0,3984	0,3981	0,4003	**0,4027**	0,4021

Table 2: German monolingual task: AvgP without query expansion using compound nouns

lack of an algorithm to split compound nouns should us to use a list of more frequently compound noun made up of 200000 terms. The scores achieved for German using the compound list were better as show the table 2.

Once, the size of passage was established for each language, the following experiment was carried out in order to study the influence of query expansion. The IR-n system uses a feedback technique to apply the query expansión. The IR-n system adds to the query the T more important term from the P more relevant passage according to [2]. Table 3 shows the scores achieved by IR-n system using the 5 more important term from the five and ten more relevant passages, and using the 10 more important term from the five and ten more relevant passages. Best results were obtained using the 10 more relevant passages, and the 10 more important term for Spanish, Italian and English, and the 5 more frequent term for German an French. This experiment shows that query expansion increase the obtained scores for all languages.

3.2 Bilingual and Multilingual task

We use three different translator in order to obtain an automatic translation of queries. The three used translator were PowerTranslator, FreeTranslator[2] and Google[3]. In multilingual task (4 languages), the queries written in English were translating to French, Spanish and German. Once, we obtained the query translation, four different experiments were carried out in order to choose the best translation. The three first ones only used a translation and the last one used the merge of all translations as query. Table 3 shows the scores achieved in the four experiments developed using every document collection as the same way as monolingual task. Best scores were achieved using the merge of translations. The IR-n system was run obtained three different rank document collections in multilingual task. According to [2], there are a few simple ways to merge ranked list of documents from different collections. We use two different method: M1) the first method is to normalize the relevance score for each topic, dividing all relevance scores by the relevance score of the top most ranked document for the same topic. M2) This method uses the following formula

[2]www.freetranslation.com

[3]www.google.com/language_tools?hl=es

T		**5**		**10**	
P	**No expansion**	**5**	**10**	**5**	**10**
Spanish	0,5042	0.5176	0,5122	0,5327	**0,5441**
Italian	0,4207	0,4428	0,4583	0,4491	**0,4679**
German	0,4027	0,4379	**0,4499**	0,4148	0,4438
French	0,4731	0,4991	**0,5286**	0,4980	0,5114
English	0,5057	0,5108	0,5034	0,5066	**0,5139**

Table 3: AvgP Using question expansion

Traslation		Free	Power	Babel	Power+Free+Babel
Spanish	0,5042	0.4235	0.4336	0.4217	**0.4371**
Italian	0,4207	0.3367	0.3490	0.3480	**0.3663**
German	0,4027	0.3037	0.3092	0.3024	**0.3245**
French	0,4731	0.3835	0.4281	0.4077	**0.4291**

Table 4: Translation used as monolingual task

		Precision at N documents					
	Cob.	5	10	20	30	200	AvgP
M1	61.97	0.6760	0.6360	0.5860	0.5367	0.2755	0.3108
M2	72.42	0.6760	0.6480	0.6030	0.5653	0.3152	0.3621

Table 5: Scores achieved by IR-n system with document merging

to normalize the document.

$$rsv'_j = (rsv_j - rsv_{min})/(rsv_{max} - rsv_{min}) \quad (5)$$

in which rsv_j is the original retrieval status value, and rsv_{min} and rsv_{max} are the minimum and maximum document scores values that a collection could achieve for the current request. Table 5 shows the scores achieved using both merging method. These scores show that best results are obtained using M2 merging method.

In bilingual task, an additional problem was found. We do not have a direct translator Italian-Spanish and Spanish-Italian. We had to translate Italian to English and late English to Spanish. This process carries out more errors than a directly translation. Table 6 shows the scores achieved in the bilingual task. At the same way as multilingual task, the best score was obtained using the merge of translations.

4 Evaluation at Clef-2003

The IR-n system used in order to participate in CLEF'2003 was the best IR-n configuration obtained in the training process using the CLEF'2002 collection.

The following subsections show the runs carried out and the scores achieved in the monolingual, bilingual and multilingual tasks.

4.1 Monolingual task

Two different runs have been submitted for each Spanish, French and Italian monolingual tasks. The first run does not use query expansion and the last one using it (IRn-*xx*-noexp and IRn-*xx*-exp, where *xx* are the language –*es*, *fr* or *it*–). Four different runs were submitted for German. The first and second runs follows the same strategies as previous languages but without splitting the compound nouns (IRn-*al*-noexp-nsp and IRn-*al*-exp-nsp). The third and fourth experiments used the splitting of compound nouns with and without expansion (IRn-*al*-noexp-sp and IRn-*al*-exp-sp)

Tables 7, 8 , 9 and 10 show the scores achieved for each run in the monolingual task. IR-n system using query expansion obtained better results than the average scores of CLEF 2003 systems for Spanish, French and German and lower scores in Italian.

Traslation		Free	Power	Babel	Power+Free+Babel
Italian-Spanish	0,4207	0.3367	0.3490	0.3480	**0.3663**

Table 6: Bilingual scores using question expansion

	AvgP	Dif.
Clef Average	0.4649	
IRn-es-exp	0.5056	**+8.75%**
IRn-es-noexp	0.4582	-1.44%

Table 7: CLEF-2003 official results. Spanish monolingual task.

	AvgP	Dif.
Clef Average	0.4843	
IRn-fr-exp	0.5128	**+5.88%**
IRn-fr-noexp	0.4853	0%

Table 8: CLEF-2003 official results. French monolingual task.

	AvgP	Dif.
Clef Average	0.4903	
IRn-it-exp	0.4802	-2.06%
IRn-it-noexp	0.4547	-7.26%

Table 9: CLEF-2003 official results. Italian monolingual task.

	AvgP	Dif.
Clef Average	0.4759	
IRn-al-nexp-nsp	0.4267	-10.34%
IRn-al-exp-nsp.	0.4687	-1.51%
IRn-al-nexp-sp	0.4670	-1.87%
IRn-al-exp-sp.	0.5115	**+7.48%**

Table 10: CLEF-2003 official results. German monolingual task.

	AvgP	Dif.
Clef Average	0.3665	
IRn-ites-noexp	0.3662	0%
IRn-ites-exp	0.4610	**+25.78%**

Table 11: CLEF-2003 official results. Italian-Spanish bilingual task.

	AvgP	Dif.
Clef Average	0.2752	
IRn-m-noexp-nsp	0.3024	+9.88%
IRn-m-exp-nsp.	0.3281	+19.22%
IRn-m-noexp-sp	0.3074	+11.7%
IRn-m-exp-sp.	0.3377	**+22.71%**
IRn-mi-exp-sp.	0.3373	+22.56%

Table 12: CLEF-2003 official results: Multilingual task.

4.2 Bilingual task

Two different runs have been submitted for bilingual tasks (Italian-Spanish). The first run does not use query expansion and the last one using it (IRn-*ites*-noexp and IRn-*ites*-exp). The English was used as intermediate language due to the lack of a direct translator Italian to Spanish. Table 11 shows that IR-n system using query expansion for bilingual task increase around a 26% the average scores of CLEF'2003 bilingual system.

4.3 Multilingual task

Five runs were submitted to Multilingual task made up four languages. The first runs (IRn-m-noexp-nsp) shows the scores achieved by IR-n system without query expansion and without splitting the German compound nouns. The second one (IRn-m-exp-nsp) presents the performance of IR-n system using query expansion and without splitting the compound nouns. The third and fourth runs (IRn-m-noexp-sp and IRn-m-exp-sp, respectively) are the same experiments but using the splitting of German compound noun. Finally, an additional experiment (IRn-mi-exp-sp) was carried out using the same passage's size for all languages (10 sentences), and using the query expansion and the splitting of compound nouns. This size was obtained experimentally in the training task.

Table 12 shows that IR-n system improve the average scores of CLEF'2003 around a 23% using the AvgP measure. Moreover, IR-n system also obtains around a 23% of improvement using he same size of passages.

5 Conclusions

General conclusions are positive. On the one hand, IR-n system has obtained better results than the average of CLEF'2003, excluding for Italian monolingual task. Moreover, we want to remark that all runs submitted are carried out only using short queries (title and description) and the average provided by CLEF organization is made up for all system (systems using both short or long queries). On the other hand, the achieved improvement using a list of the most frequent compound nouns in German conduct us to develop for next participation an algorithm to split the compound nouns.

Also, we want to emphasize the good performance of IR-n system in our first participation in bilingual and multilingual tasks. However, we planned to use a new method but time problem stopped us to submitted a new run. We hope to participate with the new method in the next conference.

Moreover, we want to underline the good scores achieved using the same size of passages for all languages. Finally, we want to emphasize that IR-n system is an information retrieval system based on passages and independent of languages according to the scores obtained in our participation at CLEF 2003.

References

[1] James P. Callan. Passage-Level Evidence in Document Retrieval. In *Proceedings of the 17th Annual International Conference on Research and Development in Information Retrieval*, pages 302–310, London, UK, July 1994. Springer Verlag.

[2] Aitao Chen. Cross-Languaje Retrieval Experiments at CLEF-2002. In CLEF [4], pages 5–20.

[3] J. Chen, A. Diekema, M. Taffet, N. McCracken, N. Ozgencil, O. Yilmazel, and E. Liddy. Question Answering: CNLP at the TREC-10 Question Answering Track. In *Tenth Text REtrieval Conference (Notebook)*, volume 500-250 of *NIST Special Publication*, Gaithersburg, USA, nov 2001. National Institute of Standards and Technology.

[4] *Workshop of Cross-Language Evaluation Forum (CLEF 2002)*, Lecture notes in Computer Science, Roma, Italy, 2002. Springer-Verlag.

[5] M. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *Sixteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68, Pittsburgh, PA, jun 1993.

[6] Marcin Kaszkiel and Justin Zobel. Effective Ranking with Arbitrary Passages. *Journal of the American Society for Information Science (JASIS)*, 52(4):344–364, February 2001.

[7] Fernando Llopis. *IR-n un sistema de Recuperación de Información basado en pasajes*. PhD thesis, Universidad de Alicante, 1998.

[8] Fernando Llopis and José L. Vicedo. IR-n system, a passage retrieval system at CLEF 2001. In *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, pages 244–252, Darmstadt, Germany, 2001. Springer-Verlag.

[9] Fernando Llopis and José L. Vicedo. IR-n system at CLEF 2002. In CLEF [4], pages 169–176.

[10] R. Muñoz and M. Palomar. *Emerging Technologies in Accounting and Finance*, chapter Sentence Boundary and Named Entity Recognition in EXIT System: Information Extraction System of Notarial Texts, pages 129–142. 1999.

[11] Gerard Salton and Chris Buckley. A term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–123, 1988.