IR-n r2 : Using normalized passages

Fernando Llopis and Rafael Muñoz and Rafael M. Terol and Elisa Noguera
Grupo de investigación en Procesamiento del Lenguaje y Sistemas de Información.
University of Alicante, Spain
{*llopis,rafael,rafamt,elisa*}*@dlsi.ua.es*

Abstract

This paper describes the fourth participation of IR-n system (Alicante University) at CLEF conferences. At present conference, we have modified the similarity measure and the query expansion model. Concerning the similarity measure, we use the normalization based on the number of words for each one of the passages. Finally, we test two different approaches for query expansion: the first one is based on documents and the second one is based on passages.

1 Introduction

In the line showed at previous conferences, IR-n system will participate in several tasks of CLEF'2004. We exactly will participate in monolingual, bilingual and multilingual tasks. IR-n system has considerably changed. On the one hand, the IR-n system was re-programmed in order to improve the answer speed. Moreover, the new version of IR-n system can use different similarity measures by means of a parameter. On the other hand, several changes were made in order to improve the similarity measures and the query expansion.

This paper is organized as follows: next section describes IR-n system and its new changes. Following, we describe the task developed at CLEF 2004 by our system. And finally, we present the achieved results and the conclusions.

2 IR-n system

Information Retrieval (IR) systems have to find the relevant documents to an user query from a document collection. We can find different kinds of IR systems at the literature. On the one hand, if the document collection and the user query are written in the same language then the IR system can be defined like a monolingual IR system. On the other hand, if the document collection and the user query are written in different languages then the IR system can be defined like a bilingual (two different languages) or multilingual (more than two languages) IR system. Obviously, the document collection for multilingual systems is written in two different languages at least. IR-n system is a monolingual, bilingual and multilingual IR system based on passages.

Passage Retrieval (PR) systems are information retrieval systems that determine the similarity of a document with regard to a user query according to the similarity of fragments of the document (passages) with regard to the same query.

There are a lot of proposals [1, 7] in order to define the best way of obtaining the passages for achieving the better results.

2.1 IR-n system r1 (2000-2003)

IR-n system was originally developed using the C++ language programming and running in linux without excessive requirements. IR-n system is a PR system that uses the sentences as atoms with the aim to define the passages. Thus each passage is composed by a specific number of

sentences. This number depends in a great measure of the collection used. For this reason, the system requires a training phase to improve its results. IR-n system uses overlapping passages in order to avoid that some documents can be considered not relevant if it appears words of the question in adjacent passages.

From the beginning, IR-n system used the traditional cosine measure [14]. However, further experiments were performed using other similarity measures which results were better than the previous ones. The similarity measures used by IR-n system differ from traditional IR systems. For example, IR-n does not use 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, IR-n system calculates the similarity between a passage P and the user 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 term t in the passage or in the question Y. N is the total number of documents in the collection, and $freq_t$ is the number of different documents that contain term t.

Once the system caculates this score for each one of the passages, it is necessary to determine the similarity of the document that contains these passages. All PR systems calculate the similarity measure of the document according to the similarity measure of their passages using the sum of similarity measures for each passage or using the best passage similarity measures for each one of the documents. The experiments performed in [6] have been re-run by IR-n system, obtaining better results when the use of the best passage similarity measures was performed as the similarity measure of the document.

Our approach 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 passage similarity measure then the system has the same behaviour as a document-based IR system adding proximity concepts.

Moreover, the use of the best passage similarity measure offers the possibility to retrieve the best passage, thus further improving the search process.

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

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

According to most of IR systems, IR-n system uses also techniques of query expansion. Originally, first release of IR-n system [9] incorporated synonyms to the original query obtaining worse scores that the model without query expansion. After that, we incorporated the model proposed in [3], but the terms that were added to the original question were the most frequent terms of the most relevant passages instead of the most frequent terms of the most relevant documents. The use of these techniques permitted us to improve our results practically in all the performed tests.

2.2 IR-n system r2 (2004)

A set of changes have been developed in our system in order to improve it. These changes are the following:

1. Firstly, we have modified the similarity measure in order to take in account the size of the passages in addition to the number of sentences used in the first release (IR-n r1). For each word, IR-n r1 stored the document and the sentences in which it was found, but did not

store the size of each one of the sentences. In this way, it was not possible to compare the similarities between passages using the size of passages. This fact has supposed an important change in the index task and in the search process. We did some experiments with pivoted cosine and okapi measures. We obtained better results with okapi measure.

2. Moreover, the system was updated in order to consider different similarity measures such as Okapi system. In this way, we can test the best setup for each document collection.

3. This new release applies techniques of query expansion based on documents. The first release of IR-n system used the most frequent terms in the passages to add them to the original question. The new IR-n release presents a new approach based on adding the most frequent terms in the documents instead of passages.

4. One of the most important factor for an information retrieval system is the speed. The first release of IR-n system had a low answer time, nevertheless the system has a delayed time in writing the most relevant passages. This fact caused that if the system used query expansion then the answer time will increase.

In order to obtain these objectives, we have affronted the decision to develop IR-n system r2 practically from the beginning. We decided to use a object oriented approach using C++ as language of implementation. Moreover in parallel way, a IR-n r3 has been developed using the ".net" technology [5]. Nowadays this release is found in an early phase of development, but it has just achieved better results for XML documents.

Also, this year we developed the web searcher of University of Alicante using IR-n system. You can access it from the web of University of Alicante (www.ua.es) or directly (www.tabarca.com).

3 IR-n r2 at Clef-2004

This year our system will participate in the following tasks:

- monolingual tasks:
  - French
  - Portuguese
  - Finnish
  - Russian
- bilingual tasks:
  - Spanish-Russian
  - English-Russian
  - Spanish-Portuguese
  - English-Portuguese
- multilingual tasks:
  - English-French-Finnish-Russian

3.1 Monolingual Tasks

We used the main resources available in the web address http://www.unine.ch/info/clef/. We take from this website stemmers and stop-word list for each language. Moreover, we used the program to convert Cyrillic characters into ASCII characters in order to process Russian documents.

Table 1: AvgP without query expansion

	Passage size using number of sentences	
Similarity measure	**Normalize okapi**	**no normalized okapi**
Finnish	0.4968	0.5011
Russian	0.4179	0.4180
French	0.4263	0.4939
English	0.5233	0.4827

Nevertheless at the moment to perform the Portuguese task, it was not available a Portuguese stemmer. For this reason, we decide to develop one. We changed the Spanish terminations using the adequate Portuguese ones.

Finnish language presents an additional feature, compound noun. A compound noun usually is composed by the combination of two or more free elements, which are morphemes that can stand on their own and that have their own meaning but together form another word with a modified meaning. We develop an algorithm for splitting compound noun into several words. This fact permitted us to improve the results in the training phase. According to [8], the split process consists on splitting words over 6 letters into known words. Obviously, we can split a word in different ways. For this reason, we use a frequency list extracted from the same corpus. We choose the known words combination that provide the highest frequency with a minimum number of words using the following formula:

$$argmin_S(\prod_{p_i \epsilon S} count(p_i)))^{1/n} \tag{5}$$

The similarity measure used for all languages in monolingual task was okapi measure obtaining the best scores. We developed several test with and without normalization using the passage size. Different scores were achieved according to the language as shows Table 1.

Similar scores were achieved for Finnish and Russian languages with the possibility of using normalization. However, for English language the system achieved best scores using a normalized measures and for French language the best scores were achieved without normalization. This is due to the fact that we have not chosen the same parameters for okapi system or well that in this case is preferably to use other similarity measure.

Different tests were performed in order to add the best approach to query expansion. Moreover, for each test we checked the adequate number of passages and documents should be considered using 5 or 10 words and 5 and 10 documents/passages. The results obtained in the experimentation phase were similar to them obtained in the final tests.

3.2 Bilingual Task

The participation of the system IR-n r2 in the bilingual task this year has been focused on the following language pairs:

- English-Russian
- Spanish-Russian
- English-Portuguese and
- Spanish-Portuguese.

According the strategy used last year by IR-n r1, the bilingual task has been performed merging several translation built by an on-line translator. This strategy is based on the idea that the words that appears in different translations have more relevancy that those that only appear in one translation.

Table 2: CLEF 2004 official results: Monolingual tasks.

Language	Run	AvgP	Dif.
Russian	CLEF Average	0.3700	
	nexp	0.4809	**+29.97%**
	pexp	0.4733	
	dexp	0.4796	
French	CLEF Average	0.4370	
	nexp	0.4086	
	pexp	0.4251	**-2.72%**
	dexp	0.4217	
Finish	CLEF Average	0.5096	
	nexp	0.4533	
	pexp	0.4908	
	dexp	0.4914	**-3.57%**
Portuguese	CLEF Average	0.4024	
	nexp	0.3532	
	pexp	0.3750	
	dexp	0.3909	**-2.85%**

Two translators were used for all languages: Freetanslation[1] and Babel Fish[2]. An additional on-line translator was used for Russian language. This translator was IMTranslator[3]. Freetranslator and Babel Fish have not a direct translation for Spanish to Russian for this reason we used English language as intermediate language.

3.3 Multilingual task

We use the formula described in [2] ir order to merge the different lists of relevant documents for each language.

$$rsv'_j = (rsv_j - rsv_{min})/(rsv_{max} - rsv_{min}) \tag{6}$$

We can not test the merging procedure in the training phase due to we conclude its implementation before to send the test results.

4 Results

4.1 Monolingual tasks

The results achieved in the monolingue task are at least peculiar. In general, all results excluding Russian results are down of the average. We considered that the results are acceptable due to our test only uses the title and the description. However, according to the training phase the Russian scores are impressive very over the average. We do not know all about the Russian language, for this reason we do not know why the results using the same release of IR are also superior to the average.

Table 2 shows the results for each language using the model without expansion (nexp), the model with expansion based on documents (dexp) and the model with expansion based on passages (pexp). The results of the best model in each case are compared with the CLEF average.

[1] www.freetranslation.com
[2] http://world.altavista.com/
[3] http://translation.paralink.com/translation.asp

Table 3: CLEF 2004 official results: Monolingual tasks.

Language	Run	AvgP	Dif.
Spanish-Rusian	CLEF Average	0.1741	
	nexp	0.3044	
	pexp	0.3087	**+77.31%**
English-Rusian	nexp	0.3296	
	exp	0.3357	**+92.82%**
	CLEF Average	0.3316	
English-Portuguese	Free-translator		
	nexp	0.2379	**-28.25%**
	pexp	0.2173	
Spanish-Portuguese	nexp	0.2977	
	exp	0.3243	**-2.2%**
English-Portuguese	Free-translator-Google-BabelFish		
	nexp	0.2975	
	pexp	0.3123	**-5.83%**

Table 4: CLEF 2004 official results: Multilingual tasks.

Language	Run	AvgP	Dif.
English	CLEF Average	0.2339	
	nexp	0.2204	
	pexp	0.2353	**+0.6%**
	dexp	0.2330	

4.2 Bilingual results

Obviously, the results achieved in bilingual tests were affected by the results of monolingual track. In this way, the English-Russian and the Spanish-Russian scores are over the average whereas the results on the English-Portuguese and Spanish-Portuguese are worse than the average. Table 3 shows the scores achieved for each pair of languages with and without query expansion.

4.3 Multilingual results

Table 4 shows the scores achieved in multilingual task without using the query expansion (nexp). Moreover, Table 4 presents the scores achieved using two different types of query expansion: the first one uses the model based on passages (pexp) and the second one uses the model based on documents (dexp). Best scores achieved for each pair of language was compared against the CLEF average as shows the *Dif.* column.

5 Conclusions and future works

This year our sensations are some contradictory. We did not have enough time to develop a new architecture system and to training it. We needed two weeks more to be able of tuning our system in order to increase the scores.

First of all, it is a surprise for us the excellent scores achieved in the bilingual task using the Russian as target language. We did not have any previous experience in this language and the system was the same for every task. For this reason, we can explain why the scores are better in Russian language than in other language.

An additional aspect to be considered it is the use of normalization. We could not demonstrate that the normalization improved the scores in all cases. Due to in the performed experiments this

fact has not been able to contrast. We want to run additional experiments in order to study this fact.

We are presented a comparison between two different query expansion models. Both models achieve similar scores but the model based on passages instead of documents is faster than the other one. Moreover, we want to check the efficiency of this model using larger documents than the CLEF collection.

Concerning to bilingual task and according our experience of last year, we follow achieving the best scores using a merging of different translator than using only all the translations. Another conclusions is that English is the best language origin to use (bilingual tests on Russian) at least if the both languages are very different. Quite the opposite occurs if both language have the same root: Romanic, Slavian, etc. (bilingual tests Spanish - Portuguese).

Concerning the multilingual task, the achieved results have been worse than the last year although they are slightly on the average of the CLEF systems. We already have known that the used model to merge document list is very dependent of the number of questions that have answer in each language. For the next edition we will hope to count on a new merging model which we have not been able to finish in this edition. Due to the results obtained int the first experiments we think this model will improve our results.

Finally, we want remark that the spent time on developing the IR-n r2 system has been allow us to do updates easily. In addition we want to emphasize the process speed in order to show the relevant passages.

6 Acknowlegments

This work has been partially supported by the Spanish Government (CICYT) with grant TIC2003-07158-C04-01.

References

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

[2] Chen, A.: Cross-Language Retrieval Experiments at CLEF-2002. In Peters et al. [4], 5–20.

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

[4] Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.): Proceedings of the Cross-Language Evaluation Forum (CLEF 2002)., Lecture Notes in Computer Science, LNCS 2785. Springer-Verlag 2003.

[5] García-PuigCerver, H., Llopis, F., Cano, M., Toral, A., Espí, H.: IR-n system, a Passage Retrieval Architecture. Proceedings of Text Speech and Dialogue 2004. Brno, September 2004

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

[7] Kaszkiel, M., Zobel, J.: Effective Ranking with Arbitrary Passages. Journal of the American Society for Information Science and Technology (JASIST), 52(4)(2001) 344–364.

[8] Koehn, P., Knight, K.: Empirical Methods for Compond Splitting. Proceding of EACL 2003.

[9] Llopis, F., Vicedo, J.L.: IR-n system at CLEF 2002. In Peters et al. [4], 169–176.

[10] Llopis, F.: IR-n un sistema de Recuperación de Información basado en pasajes. PhD thesis. Universidad de Alicante (2003).

[11] Llopis, F. and Vicedo, J.L.: IR-n system, a passage retrieval system at CLEF 2001. In Proceedings of the Cross-Language Evaluation Forum (CLEF 2001). LNCS 2406. Springer-Verlag (2002) 244–252.

[12] Muñoz, R. and Palomar, M.: Sentence Boundary and Named Entity Recognition in EXIT System: Information Extraction System of Notarial Texts. In Emerging Technologies in Accounting and Finance (1999) 129–142.

[13] Lernout & Hauspie Educational Software Power Translator. Software.

[14] Salton, G., Buckley, C.: Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24(5) (1988) 513–123.