# Passage Retrieval vs. Document Retrieval in the Monolingual Task with the IR-n system

Elisa Noguera and Fernando Llopis
Grupo de investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
elisa,llopis@dlsi.ua.es

**Abstract**

The paper describes our participation in monolingual tasks at CLEF 2006. We have submitted results for the following languages: English, French, Portuguese and Hungarian. We focused on studying different weighting schemes (okapi and dfr) and retrieval strategies (passage retrieval and document retrieval) to improve retrieval performance. After an analysis of our experiments and of the official results at CLEF, we find that our different configurations (French, Portuguese and Hungarian) achieve considerably improved scores.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval

## General Terms

Experimentation, Measurement, Performance

## Keywords

Information Retrieval

## 1 Introduction

In our sixth participation at CLEF, we focused on evaluating: a new weighting model (dfr), retrieval based on passages/documents and setting the best configuration to each language. Specifically, we participated in the following languages: English, French, Portuguese and Hungarian.

IR-n system [5] was developed in 2001. It is a Passage Retrieval (PR) system which uses passages with a fixed number of sentences. This provides the passages with some syntactical content. Previous researches with the IR-n system ([6] [7] [8] [9]) are based on detecting the suitable size for each collection ( to experiment with test collection ), but determining the similarity of a document based on the passage with more similarity. Last year [10] we proposed a new method 'combined size passages' in order to improve the performance of the system combining different size passages. This year we implemented a new similarity measure in the system and we tested our system with different configurations to each language.

Futhermore, our team participated in other tasks at CLEF-2006 as GeoCLEF, CL-SR... and we also applied PR systems in other tasks, such as Question Answering (QA) [3].

This paper is organized as follows: next section describes IR-n system and its new changes. Following, we describe the task developed at CLEF 2006 by our system and the training. And finally, we present the achieved results and the conclusions.

# 2 IR-n system

In this section the main characteristics of the IR-n system are presented and details are given on the resources and the techniques of the system used in CLEF 2006.

## 2.1 Resources: stemmers and stopword lists

We used the stemmers and stopwords lists available on the web http://www.unine.ch/info/clef. We highlight that Hungarian collections are encoded in UTF-8.

## 2.2 Weighting models

IR-n system uses several similarity measures. This year, the weighting model dfr [1] was included, but we also used okapi weighting model[4].

The document ranking produced by each weighting model is represented using the same general expression, namely as the product of a document-based term weight by a query-based term weight:

$$sim(q,d) = \sum_{t \in q \wedge p} w_{t,p} \cdot w_{t,q} \tag{1}$$

**List of variables** Here is described the list of variables used in the following formulas. 2, 3:

- $f_{t,p}$ is the frequency of the term $t$ on the passage $p$,
- $f_{t,q}$ is the frequency of the term $t$ on the query $q$,
- $n$ is the number of documents in the collection,
- $n_t$ is the number of documents which $t$ appears,
- $c$, $k_1$, $b$ and $k_3$ are constant values,
- $ld$ is the length of the document,
- $avg_{ld}$ is the average of the length of the documents

**Okapi** Using the okapi model, the relevance score of a passage p for query q is given by:

$$\begin{aligned} w_{t,p} &= \frac{(k_1+1)\cdot f_{t,p}}{K \cdot f_{t,p}} \\ w_{t,q} &= \frac{(k_3+1)\cdot f_{t,q}}{k_3 \cdot f_{t,q}} \cdot w_t \\ K &= (1-b) + b \cdot \frac{ld}{avr_l d} \\ w_t &= log_2 \frac{n - n_t + 0.5}{n_t + 0.5} \end{aligned} \tag{2}$$

**DFR** Using this model, the weight of a passage p for query q is given by:

$$\begin{aligned} w_{t,q} &= f_{t,q} \\ w_{t,p} = (log_2(1+w_t) + w'_{t,p} \cdot log_2(\frac{1+w_t}{w_t})) \cdot \frac{f_t+1}{n_t \cdot (w'_{t,p}+1)} \\ w'_{t,p} &= f_{t,p} \cdot log_2(1+\frac{c \cdot avr_{ld}}{ld}) \\ w_t &= \frac{f_t}{n} \end{aligned} \tag{3}$$

### 2.3 Query expansion

Most IR systems use query expansion techniques [2] based on adding the most frequent terms contained in the most relevant documents to the original query. The IR-n architecture allows us to use query expansion based on either the most relevant passages or the most relevant documents. In previous researches, we obtained better results using the most relevant passages.

## 3 Training

This section describes the training process which has been carried out in order to obtain the best features to improve the performance of the system. Firstly, the collections and resources are described. The following section explains the specific experiments which we have carried out.

### 3.1 Data Collections

This year our system has participated in the following monolingual tasks: English, French, Portuguese and Hungarian. Table 1 shows the characteristics of the language collections.

| Language | Collections | NDocs | Size | SDAvg | WDAvg | WSAvg |
|---|---|---|---|---|---|---|
| English | The Angeles Times 94<br>Glasgow Herald 95 | 169477 | 579 MB | 25 | 529 | 20 |
| French | Le Monde 94/95<br>SDA French 94/95 | 177452 | 487 MB | 17 | 388 | 21 |
| Portuguese | Público 94/95<br>Folha 94/95 | 210734 | 564 MB | 18 | 433 | 23 |
| Hungarian | Magyar Hirlap 02 | 49530 | 105 MB | 11 | 245 | 20 |

Table 1: Data Collections

- SDAvg is the average of sentences in each document.
- WDAvg is the average of words in each document.
- WSAvg is the average of words in each sentence.

### 3.2 Experiments

The aim of the experiment phase is set up the optimum value of the input parameters for each collection. For training has been used the collections CLEF-2005 (English, French, Portuguese and Hungarian). Query expansion techniques have also been used in all languages. In addition, we describe the input parameter of the system:

- **Size passage (sp)**: We established two size passage: **8** (normal passage) or **30** (big passage).
- **Weighting model (wm)**: We use two weighting models: **okapi** and **dfr**.
- **Opaki parameters**: these are $k_1$, $b$ and $avg_{ld}$ ($k_3$ is fixed as 1000).
- **Dfr parameters**: these are $c$ and $avg_{ld}$.
- **Query expansion parameters**: If **exp** has value 1, this denotes we use relevance feedback based on passages in this experiment. But, if **exp** has value 2, the relevance feedback is based on documents. Moreover, **np** and **nd** denote the k terms extracted from the best ranked passages (np) or documents (nd) from the original query.
- **Evaluation measure**: Mean average precision (**avgP**) is the evaluation measure used in order to evaluate the experiments.

#### 3.2.1 English

As we can see at table 2, the best weighting scheme is dfr. Therefore, the passage size 8 obtains 0.5403 as average precision.

| sp | wm | c | avgld | k1 | b | exp | np | nd | avgP |
|---|---|---|---|---|---|---|---|---|---|
| 8 | dfr | 4 | 600 | | | | | | 0.4940 |
| 30 | dfr | 4 | 600 | | | | | | 0.5063 |
| **8** | **dfr** | **4** | **600** | | | **2** | **10** | **10** | **0.5403** |
| 30 | dfr | 4 | 600 | | | 1 | 5 | 10 | 0.5384 |

Table 2: Training results English 2005

#### 3.2.2 French

For French language, the best weighting scheme is okapi with 9 as passage size. This configuration has obtained 0.3701 as average precision.

| sp | wm | c | avgld | k1 | b | exp | np | nd | avgP |
|---|---|---|---|---|---|---|---|---|---|
| 8 | dfr | 3 | 300 | | | | | | 0.3090 |
| 30 | dfr | 3 | 300 | | | | | | 0.3105 |
| 50 | dfr | 2 | 300 | | | | | | 0.3102 |
| 8 | dfr | 2 | 300 | | | 2 | 10 | 10 | 0.3686 |
| 30 | dfr | 2 | 300 | | | 2 | 5 | 10 | 0.3607 |
| 9 | okapi | | 300 | 1.2 | 0.3 | | | | 0.2964 |
| 8 | okapi | | 300 | 1.2 | 0.3 | | | | 0.2954 |
| 9 | okapi | | 300 | 1.5 | 0.3 | | | | 0.3011 |
| **9** | **okapi** | | **300** | **1.5** | **0.3** | **1** | **5** | **10** | **0.3701** |
| 30 | okapi | | 300 | 1.5 | 0.3 | 1 | 5 | 10 | 0.3603 |

Table 3: Training results French 2005

#### 3.2.3 Hungarian

The best weighting scheme is dfr to Hungarian language, whereas the passage size is 30 to Hungarian. The best average precision obtained by this configuration is 0.3644.

| sp | wm | c | avgld | k1 | b | exp | np | nd | avgP |
|---|---|---|---|---|---|---|---|---|---|
| 8 | dfr | 5 | 300 | | | | | | 0.3119 |
| 30 | dfr | 2 | 300 | | | | | | 0.3333 |
| 50 | dfr | 2 | 300 | | | | | | 0.3334 |
| 8 | dfr | 5 | 300 | | | 1 | 10 | 10 | 0.3534 |
| **30** | **dfr** | **2** | **300** | | | **2** | **10** | **10** | **0.3644** |
| 8 | okapi | | 100 | 1.5 | 0.3 | | | | 0.2930 |
| 8 | okapi | | 300 | 1.2 | 0.3 | 1 | 10 | 10 | 0.3264 |

Table 4: Training results Hungarian 2005

#### 3.2.4 Portuguese

The best configuration by Portuguese language is the same as Hungarian language (dfr as weighting scheme and 30 as size passage). The average precision obtained with this configuration is 0.3948.

| sp | wm | c | avgld | k1 | b | exp | np | nd | avgP |
|---|---|---|---|---|---|---|---|---|---|
| 8 | dfr | 6 | 300 | | | | | | 0.3362 |
| 30 | dfr | 4 | 300 | | | | | | 0.3484 |
| 30 | dfr | 6 | 300 | | | | | | 0.3457 |
| 50 | dfr | 4 | 300 | | | | | | 0.3474 |
| 8 | dfr | 6 | 300 | | | 1 | 10 | 10 | 0.3733 |
| **30** | **dfr** | **4** | **300** | | | **1** | **5** | **10** | **0.3948** |
| 8 | okapi | | 300 | 1,5 | 0,3 | | | | 0.3283 |
| 8 | okapi | | 300 | 1,5 | 0,3 | 1 | 10 | 10 | 0.3676 |
| 30 | okapi | | 300 | 1,5 | 0,3 | 1 | 10 | 10 | 0.3793 |

Table 5: Training results Portuguese 2005

#### 3.2.5 Experiments summary

In conclusion, the table 6 shows the best configuration for each language. These configurations were used at CLEF 2006.

| language | run | sp | wm | C | avgld | k1 | b | exp | np | nd | avgP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **English** | 8-dfr-exp | 8 | dfr | 4 | 600 | | | 2 | 10 | 10 | 0.5403 |
| **French** | 9-okapi-exp | 9 | okapi | | 300 | 1.5 | 0.3 | 1 | 5 | 10 | 0.3701 |
| **Hungarian** | 30-dfr-exp | 30 | dfr | 2 | 300 | | | 2 | 10 | 10 | 0.3644 |
| **Portuguese** | 30-dfr-exp | 30 | dfr | 4 | 300 | | | 1 | 5 | 10 | 0.3948 |

Table 6: Configurations used at CLEF 2006

## 4 Results at CLEF-2006

We submitted four runs for each language in our participation (except for English that we have submitted 1 run) in CLEF 2006. The best parameters, i.e. those that gave the best results in system training, were used in all cases.

This is the description of the runs that we submitted at CLEF 2006:

- yy-xx-zexp
  - yy is the passage size
  - xx is the weighting model used (dfr or okapi)
  - z is the expansion query (not used 'nexp')

The official results for each run are showed in Table 7. Like other systems which use query expansion techniques, these models also improve performance with respect to the base system. Our results are appreciably above average in all languages, except for English where they are sensible below the average. This results present that the percentage of improvement in Portuguese is 15.43% in avgP.

| **Language** | **Run** | **AvgP** | **Dif** |
|---|---|---|---|
| English | CLEF Average | 38.73 | |
| | 30-dfr-exp | 38.17 | -1.44% |
| French | CLEF Average | 37.09 | |
| | 30-dfr-exp | 37.13 | |
| | 8-dfr-exp | 38.28 | +3.2% |
| | 9-okapi-exp | 35.28 | |
| | 30-okapi-exp | 37.80 | |
| Portuguese | CLEF Average | 37.32 | |
| | 30-dfr-exp | 41.95 | |
| | 8-dfr-exp | 42.06 | |
| | 8-okapi-exp | 42.41 | |
| | 30-okapi-exp | 43.08 | +15.43% |
| Hungarian | CLEF Average | 33.37 | |
| | 30-dfr-exp | 35.32 | +5.5% |
| | 8-dfr-exp | 34.25 | |
| | 30-dfr-nexp | 30.60 | |
| | 8-dfr-nexp | 29.50 | |

Table 7: CLEF 2006 official results. Monolingual tasks.

# 5 Conclusions and Future Work

In this seventh CLEF evaluation campaign, we proposed a different configuration for the English, French, Portuguese and Hungarian languages (see table 6). In order to enhance retrieval performance, we have evaluated different weighting models using also a query expansion approach based on passages and documents.

The results of this evaluation indicate that for the French, Portuguese and Hungarian languages proved to be effective (see table 7) because the results are above average. However, the English language has obtained results sensible below average.

For Portuguese language, the best results are obtained by okapi weighting model. For other languages (English, French and Hungarian), the best results are obtained by dfr (see table 7).

The best passage size for French was 9, although for other languages (English, Portuguese and Hungarian) was 30 (this passage size is comparable to IR based on the complete document).

As in previous evaluation campaigns, pseudo-relevance feedback based on passages improves mean average precision statistics for all languages, even though this improvement is not always statistically significant.

Lastly, we outline the future directions that we plan to undertake are evaluate languages as Bulgarian or Spanish. Therefore, as future work we also consider to research into different ways of providing Natural Language information to basic IR and evaluating the impact of each approach.

## 6 Acknowledgements

This research has been partially funded by the Spanish Government under project CICyT number TIC2003-07158-C04-01 and by the Valencia Government under project number GV06-161.

## References

[1] G. Amati and C. J. Van Rijsbergen. Probabilistic Models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.

[2] Aitao Chen and Fredric C. Gey. Combining Query Translation and Document Translation in Cross-Language Retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, pages 108–121, Trondheim, Norway, 2003. Springer-Verlag.

[3] Noguera Elisa, Fernando Llopis, and Antonio Ferrández. Passage Filtering for Open-Domain Question Answering. In Sampo Pyysalo Tapio Salakoski, Filip Ginter and Tapio Pahikkala, editors, *Advances in Natural Language Processing, Proceedings of 5th International Conference on Natural Language Processing FinTAL*, volume 4139 of *Lecture Notes in Computer Science*, pages 756–767. Springer-Verlag, August 2006.

[4] Savoy J. Fusion of Probabilistic Models for Effective Monolingual Retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, Trondheim, Norway, 2003. Springer-Verlag.

[5] F. Llopis. *IR-n: Un Sistema de Recuperación de Información Basado en Pasajes*. PhD thesis, University of Alicante, 2003.

[6] Fernando Llopis and José Luis Vicedo González. IR-n: A Passage Retrieval System at CLEF-2001. In *CLEF*, pages 244–252, 2001.

[7] Fernando Llopis, José Luis Vicedo González, and Antonio Ferrández. IR-n System at CLEF-2002. In *Proceedings of CLEF 2002*, pages 291–300, 2002.

[8] Fernando Llopis and Rafael Muñoz. Cross-Language Experiments with the IR-n System. In *Proceedings of CLEF 2003*, 2003.

[9] Fernando Llopis, Rafael Muñoz, Rafael M. Terol, and Elisa Noguera. IR-n r2 : Using normalized passages. In Carol Peters and Francesca Borri, editors, *Cross Language Evaluation Forum: Working Notes for the CLEF 2004 Workshop*, pages 65–72, Pisa, Italy, 2004. IST-CNR.

[10] Fernando Llopis and Elisa Noguera. Combining Passages in the Monolingual Task with the IR-n System. In *Proceedings of CLEF 2005*, 2005.