Lecture Notes in Computer Science

The LNCS series reports state-of-the-art results in computer science research, development, and education, at a high level and in both printed and electronic form. Enjoying tight cooperation with the R&D community, with numerous individuals, as well as with prestigious organizations and societies, LNCS has grown into the most comprehensive computer science research forum available.

The scope of LNCS, including its subseries LNAI, spans the whole range of computer science and information technology including interdisciplinary topics in a variety of application fields. The type of material published traditionally includes

- proceedings (published in time for the respective conference)
- post-proceedings (consisting of thoroughly revised final full papers)
- research monographs (which may be based on outstanding PhD work, research projects, technical reports, etc.)

More recently, several color-cover sublines have been added featuring, beyond a collection of papers, various additional components. These sublines include

- tutorials (textbook-like monographs or collections of lectures given at advanced courses)
- state-of-the-art surveys (offering complete and mediated coverage of a topic)
- hot topics (introducing emergent topics to the broader community)

In parallel to the printed book, each new volume is published electronically in LNCS Online at *http://www.springerlink.com/series/lncs/*

Detailed information on LNCS can be found at the series home page *http://www.springer.de/comp/lncs/*

Proposals for publication should be sent to

LNCS Editorial, Tiergartenstr. 17, 69121 Heidelberg, Germany

E-mail: lncs@springer.de

ISSN 0302-9743

ISBN 3-540-40830-4

Lecture Notes in Computer Science

LNCS LNAI LNBI

springeronline.com

Peters · Braschler
Gonzalo Kluck (Eds.)

LNCS 2785

Advances in Cross-Language Information Retrieval

LNCS 2785

CLEF 2003

Carol Peters
Martin Braschler
Julio Gonzalo
Michael Kluck (Eds.)

Advances in Cross-Language Information Retrieval

Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002 Rome, Italy, September 2002 Revised Papers

CLEF

Springer

CLEF 2002 Workshop Steering Committee

Martin Braschler	Eurospider Information Technology AG, Switzerland
Khalid Choukri	Evaluations and Language Resources Distribution Agency, France
Julio Gonzalo Arroyo	Universidad Nacional de Educación a Distancia, Spain
Donna Harman	National Institute of Standards and Technology, USA
Noriko Kando	National Institute of Informatics, Japan
Michael Kluck	IZ Sozialwissenschaften, Bonn, Germany
Patrick Kremer	Institute de Information Scientifique et Technique, Vandoevre, France
Carol Peters	Italian National Research Council, Pisa, Italy
Peter Schäuble	Eurospider Information Technology AG, Switzerland
Laurent Schmitt	Institute de Information Scientifique et Technique, Vandoevre, France
Ellen Voorhees	National Institute of Standards and Technology, USA

Table of Contents

Monolingual Experiments

Mainly Domain-Specific Information Retrieval

Interactive Track

Cross-Language Spoken Document Retrieval

II Cross-Language Systems Evaluation Initiatives, Issues and Results

III Appendix

Universities of Alicante and Jaen at iCLEF

Fernando Llopis[1], José L. Vicedo[1], Antonio Ferrández[1], Manuel C. Díaz[2], and Fernando Martínez[2]

[1] Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
{llopis,vicedo,antonio}@dlsi.ua.es
[2] Departamento de Ciencias de la Computación
University of Jaen, Spain
{mcdiaz,dofer}@ujaen.es

Abstract. We present the results obtained at iCLEF-2002. This is the first time that we have participated in the iCLEF task, and we have used our Passage Retrieval approach (IR-n). This system previously divides the document in fragments or passages, and the similarity of each passage with the query is then measured. Finally, the document that contains the most similar passage is returned as the most relevant. In the interactive document selection task, we have experimented with this system by showing the most relevant passage of each returned document instead of the entire document. In this paper, we present the results obtained with this system, where the relevant passages have been automatically translated into Spanish by means of Systran. The results are compared with those obtained using the Z-Prise system where the entire document is read to establish relevance.

1 Introduction

The focus of this paper is the interactive document selection task. The main objective of this task is to design a system to facilitate users to find relevant documents about their information needs. Classical Information Retrieval (IR) systems use the whole document in order to determine the relevance of the document with reference to a query. The main problem with this kind of system is that they can return entire relevant documents, but they cannot locate the most relevant piece of text in the document. For example, a document about the "Biography of Felipe II" is relevant for the query "the town were Felipe II was born", but only a part of this document is relevant for the information required. In this way, when users have to determine whether a document is relevant or not, they probably have to read the entire document. A new IR proposal that tries to overcome this problem is called Passage Retrieval (PR). PR systems divide the document into pieces of text that are called passages. A similarity measure is obtained for each passage and then the document will be given the similarity value of its most relevant passage. The IR system used in this paper, called IR-n, employs the PR strategy. The IR-n system was used as an IR system in CLEF 2001 [2], and as a module in a Question Answering (QA) system in

C.A. Peters (Ed.): CLEF 2002, LNCS 2785, pp. 392–399, 2003.

TREC-10 [7], where it reduces the amount of text in which the QA system works. In this paper, the results obtained with the IR-n system for the iCLEF task are presented, with the user determining whether a document is relevant or not by reading only the most relevant passages returned by this system. These relevant passages are automatically translated into Spanish by Systran.

This paper is structured in the following way. First, an introduction of PR systems and the architecture of IR-n system are presented. Second, the interactive document selection task is introduced. Third, the experiments for tuning the IR-n system for this task are described and results achieved are analysed and compared with those obtained with the Z-Prise system, which is based on the IR approach that uses the entire document to determine the relevance to a query. Finally, the conclusions obtained from this experiment and plans for future work are presented.

2 The State of the Art in Passage Retrieval

Previous work [4] shows that PR systems can improve the precision of IR systems by from 20% to 50%. PR systems can be classified according to the way they define the passages in a document. A general classification usually quoted by researchers is that proposed in [1], where the PR systems are divided into those based on discourse, those based on semantic properties and those based on a window model. The first one uses the structural properties of the documents, such as sentences, paragraphs or HTML marks in order to define the passages. The second one divides each document in semantic pieces, according to the different topics in the document. The last one uses windows with a fixed size to form the passages. We can find a further taxonomy of window models in [4], which distinguishes between those that use the structure of the document when defining the passages, and those that do not use this kind of information. On the one hand, it seems plausible that discourse-based models are more effective since they are using the structure of the document itself. However, their greatest problem is that the results could depend on the writing style of the document author. Furthermore, this kind of model produces a very heterogeneous set of passages, with reference to the size of each passage. On the other hand, window models have the great advantage that they are simpler to create since the passages have a previously known size, whereas the other models have to support variable sized passages. However, they have the problem that as the passage can start with any word in the sentence, such passages may not be adequate for presentation to the user as the most relevant passage, since they are not logical and coherent fragments of the document.

3 IR-n System Architecture

The IR-n system [2] is based on a window model that uses the structure of the document when defining the passages. The main characteristics of this system are the following:

1. A document is divided into passages, which are formed by a fixed number of sentences. This is because a sentence usually represents an idea in the document, whereas the paragraphs can be used just to give a visual structure to the document. Moreover, sentences are logical and complete units of information, whereas window models that start with any word in the document can return incoherent fragments of text.
2. The number of sentences that form a passage can be separately determined for each set of documents. Previous experiments for the documents of the Los Angeles Times show that the best results are obtained with passages of seven sentences.
3. The system uses windows with overlapped pieces of text in order to fine-tune the results. For example, with passages of seven sentences, the first passage is formed by sentences from 1 to 7, the second one from 2 to 8, and so on. We have used these overlapping passages because we have obtained better results in the experiments presented in [5], than when using other kinds of passages (e.g. those with no overlapping, or with other degrees of overlapping). The overlapping process increments the running time, but this increment is not very high, since the first passage starts in the first sentence where a key word of the query appears, and the last passage in the last sentence where a key word appears.
4. We are using the cosine measure but with no normalization with reference to the size of the passage, because the passages are quite homogeneous (the same number of sentences with a similar number of words).

4 The Interactive Document Selection Task in CLEF-2002

The aim of the interactive document selection (IDS) task in CLEF-2002 is to evaluate the systems that allow multilingual interactive information retrieval. The idea of interactivity is related to the issues involved in displaying documents appropriately so that users can decide whether they are relevant or not. Multilinguality means that the answer is presented in one language while the document collection is written in a different one. The set-up of this task is the following:

1. Each participant has to compare two IDS systems. One of them will be taken as baseline.
2. Two four-user groups are selected by each participant.
3. Each individual user has to process a set of four queries written in their native language.
4. The document collection where the search has to be performed is written in a different language to the queries.
5. The relevant documents obtained by an IR system are shown to the user for each query. These documents are sorted by relevance.
6. Each user should decide which of these documents are relevant for him/her.

7. Each participant in the task should submit the list of documents that have been selected by the users.

The organization evaluates if the documents that have been selected by iCLEF users are really relevant or not. After that, precision, recall and F_α measures are calculated for each participant. F_α [6] is defined as follows:

$$F_\alpha = \frac{1}{\alpha/P + (1-\alpha)/R} \quad (1)$$

Where $\alpha = 0.8$ since it is considered more valuable to obtain higher precision (P) than recall (R).

5 IR-n System in the Interactive Document Selection Task

The participation of the IR-n system in this task has been based on a set of Spanish queries, and a document collection written in English, the LA Times collection.

The query collection was made up by some of the CLEF-2001 queries, specifically numbers 53, 56, 65 and 80. Moreover, there was an additional query (number 86) that was used for system training. It is important to note that we have only processed short versions (title + description) of the queries.

A number of experiments were carried out with two main aims. First, in order to present the documents to the user, we designed an HTML interface as is shown in Figure 1. This interface presented the most relevant passage (seven sentences) of each document and allowed an easy selection of relevant or non-relevant documents. In addition, it collected other useful information for the task, such as the number of the question and the name of the user.

Second, as the IR-n system does not have multilingual capabilities, and in order to facilitate the document reading process, we translated the documents into Spanish using Systran[1].

5.1 System Training

The system was trained using the query supplied by the iCLEF organisation. The objective was to adapt the process by taking into account the comments and suggestions of two non-experimented users in order to investigate the best way to show the information to the user.

The main problems detected in this experiment were the following:

1. Document translation was not good enough, and produced some unreadable pieces of text that made it difficult to understand the passage presented.

[1] http://www.systransoft.com

2. Sometimes, the passage did not include some important information, because the IR-n system started building relevant passages from the first sentence that contained a query term.
3. Even when the translation was reasonable, sometimes reading only the passage was not enough to decide if it was relevant because user lacked information about the context in which the passage was included.
4. Usually, when users did not understand these pieces of text or doubted their relevance because contextual information was lacking, they discarded them and tried to read the following piece of text.

After this first experiment, several changes were introduced to the interface in order to minimize these problems and facilitate user decisions:

1. As our system lacks of translation capabilities nothing was done to improve translation. Instead, we increased readability by presenting the passages in different lines to the user. This facilitated the comprehension of the passages, and was quite easy to achieve since IR-n performs an indexing on sentences.
2. In order to avoid missing important information, the previous sentence to the relevant passage was also presented. By including this sentence, the comprehension of the passage is highly increased.
3. In order to provide important context information, the system always shows the title of the document to which the relevant passage belongs. This way, the

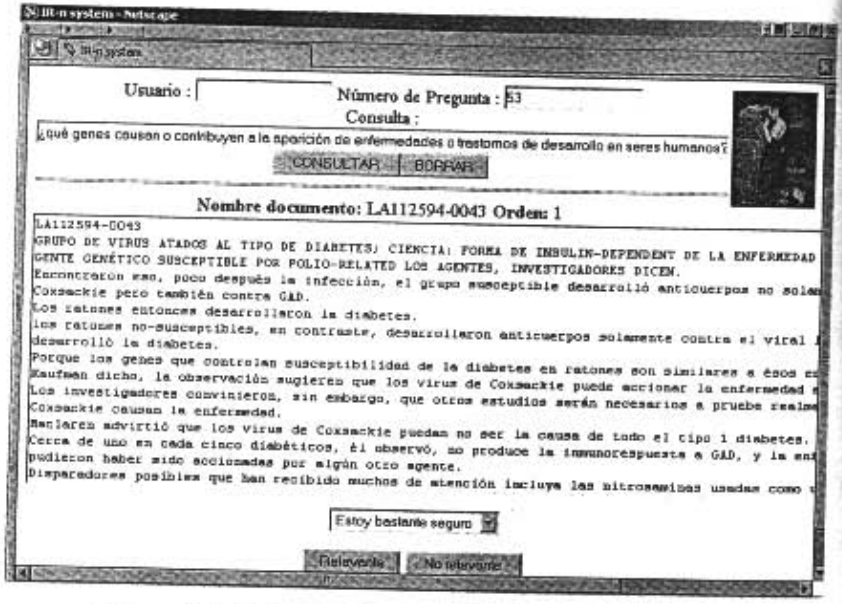

Fig. 1. HTML interface for presenting documents to the user

user knows the main subject the passage talks about and enhances passage understanding. This title is presented in capital letters.

All these changes, tested by two different users to those who performed training, produced a better readability and comprehension of passages. Moreover, the amount of discarded or skipped passages was drastically reduced.

5.2 System Evaluation

The experiments were carried out by eight final course university students. All of them processed relevant passages returned by IR-n system and full documents retrieved by the Z-Prise IR system [3] which was used as baseline.

Instead of explaining to the students what the IR community considered to be a "relevant" or a "non-relevant" document, we decided to adopt a different strategy with the aim of coordinating relevance criteria between them. They were told to select all those documents that, in their opinion, provided useful information about the topic of the query since they would later be requested to write a paper on this topic. The experiments were carried out in the following way:

- All students (users) processed all the questions in both systems (IR-n and Z-Prise).
- The 25 most relevant documents selected by the IR system were presented to the user. The IR-n system only showed the more relevant passage of each document, whereas the Z-Prise system displayed the whole document.
- The user classified each document as relevant or non-relevant.
- The time for each query was not limited.

5.3 Results

The results achieved are presented in Table 1. This table compares IR-n system results with those obtained with Z-Prise. Only the first 25 most relevant documents have been used for each query, which could explain the low recall results.

Given that only 25 relevant documents were retrieved for each topic, it is interesting to study the precision results. Table 2 presents the precision achieved for each topic. First, the low results obtained for Topic 3 should be pointed out because no relevant document was retrieved by any system. Second, we should note the high precision achieved in two of the three remaining topics, which was obtained with just the most relevant passage.

Table 1. Results comparison

System	Average F_alpha
Z-Prise	0.2166
IR-n	0.3248

Table 2. Results by topic

Average Precision	IR-n	Z-Prise
Topic 1	0.4601	0.6371
Topic 2	0.8098	0.5925
Topic 3	0	0
Topic 4	0.7643	0.3748
Average	0.5085	0.4011

6 Conclusions and Future Work

In this paper, we have described an experiment that studies the ability of users to judge the relevance of documents, in which the users can only read the most relevant passages of these documents. The results were quite good, because the users took only a short time to judge relevance since they had to read short pieces of text. However, these short pieces of text contain the most relevant data about the information required, therefore the precision results were high, even higher than those obtained by means of reading the entire document. There are some points to note, once the individual results and the opinions of the users have been analysed:

- Users become very anxious when they do not find the relevant document in the list of relevant passages. This occurred for one of the queries in which only one relevant document appeared in the 25 documents presented by our system. In this case, the users judged as relevant some non-relevant documents that were not selected in other cases.
- The users find the automatic translations into Spanish quite unreadable most of the times (more than was expected).
- We think that the results have been influenced by using just the title and description fields of the queries, which have raised some doubts about the relevance of the passages.
- It has been difficult to find users to carry out the experiments, which explains the reduction of the number of documents to study (only 25) for each query. This has highly decreased the recall of the IR-n system, although we are quite happy with the results obtained, since the users found a high percentage of relevant documents in not much time (an average between 8 and 9 minutes per query). This is because the piece of text that has to be read is only formed by seven sentences.

In the future, we plan to improve the automatic translations. Systran has been used in order to translate the passages presented to the user. Given that the automatic translation of the Los Angeles Times collection was imperfect, and even at times unreadable, we will try to present the results to the user in a more structured way. This task will be carried out by retrieving information from a collection similar to the Los Angeles Times, specifically, the EFE news of the same year, which is available in Spanish.

Secondly, we have to improve the interactivity with the system by using user relevance decisions to learn about question expansion techniques.

Acknowledgements

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.

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, London, UK, July 1994. Springer Verlag, pages 302-310.

[2] CLEF2001. Evaluation of Cross-Language Information Retrieval Systems. In Proceedings of the Workshop of the Cross-Language Evaluation Forum, Darmstadt, Germany, 2001. Lecture Notes in Computer Science 2406, Springer-Verlag.

[3] Darrin Dimmick. ZPrise information retrieval system. Available on demand at NIST. http://www.itl.nist.gov/iaui/849.02/works/papers/zp2/zp2.html.

[4] 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.

[5] Fernando Llopis, José L. Vicedo, and Antonio Ferrández. Text Segmentation for efficient Information Retrieval. In Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002), Mexico City, Mexico, 2002. Lecture Notes in Computer Science, Springer-Verlag, pages 373-380.

[6] C. J. Van Rijsbergen. Information Retrieval, 2nd edition. Butterworths, London, 1979.

[7] José Luis Vicedo, Antonio Ferrández, and Fernando Llopis. University of Alicante at TREC-10. In Tenth Text REtrieval Conference (Notebook), volume 500-250 NIST Special Publication, Gaithersburg, USA, Nov 2001. National Institute of Standards and Technology.