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. In this paper, the results obtained at iCLEF-2002 are presented. This is the first time that we try to face up the iCLEF task, and we have used our Passage Retrieval approach (IR-n). This system previously divides the document in fragments or passages, and after that, the similarity of each passage with the query is 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 work, we present the results obtained with this system, where the relevant passages have been automatically translated into Spanish by means of Systran. Moreover, the results are compared with Z-Prise system, which is based on the reading of the entire document.

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. The 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 of this kind of systems 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 a user has to determine if a document is relevant or not, he/she has to read probably the entire document. A new IR proposal that tries to overcome this problem is called Passage Retrieval (PR). The PR systems divide the document into pieces of text that are called passages. After that, the similarity measure is obtained for each passage, and finally, the document will obtain the similarity value of its most relevant passage. The IR system used in this paper, called IR-n, employs the PR strategy too. The IR-n system has been used as IR system in CLEF 2001 [2], and as a module in a Question Answering

(QA) system in TREC-10 [6], 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 iCLEF task are presented, that is to say, when the user determines if a document is relevant or not by means of 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. Firstly, an introduction of PR systems is presented. Secondly, the architecture of IR-n system and the experiments for tuning the IR-n system for the document selection task are described. Thirdly, the results are explained 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 with this work and future works are presented.

2 The state of the art in Passage Retrieval

Previous works [4] show that PR systems can improve the precision of IR systems between 20 and 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 the discourse, those based on semantic models 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. Moreover, we can find another taxonomy of window models in [4], where it is distinguished between those that use the structure of the document in the moment to define the passages, and those that do not use this kind of information. On the one hand, it seems coherent that discourse-based models are more effective since they are using the structure of the document itself. However, the greatest problem of them is that the results could depend on the writing style of the document author. Moreover, this kind of models produces a very heterogeneous set of passages, with reference to the size of each passage. On the other hand, window models have the main advantage that they are simpler to accomplish, since the passages have a previously known size, whereas the remaining models have to bear in mind the variable size of each passage. However, they have the problem that when the passage starts in whatever word of the sentence, these passages could not be adequate in order to be presented to the user as the most relevant passage, since they are not logic 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 in the moment to define 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 for giving a visual structure to the document. Moreover, the sentences are logic and complete units of information, whereas those window models that start on whatever 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 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 overlapped passages because we have obtained better results in the experiments presented in [5], than 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 Experiments

Some experiments have been carried out with two main aims: firstly, with the aim of configuring the way of presenting the documents to the user; secondly, with the aim of facilitating the reading process of the documents. Firstly, an HTML interface was created as it is shown in Figure 1.

In the HTML interface, information useful for iCLEF was collected, such as the number of the question and the name of the user. In iCLEF'2002, the most relevant passage formed by seven sentences was presented to the user. After the first experiments, several elements were added to the interface in order to facilitate the decision of the user about the relevance of each document:

1. The visualization of the first line of the title of the document in capital letters.
2. The following sentence after the title is the previous sentence to the most relevant passage. This is because in the IR-n system, the most relevant passage always starts in the first sentence where a key word of the query appears. In this way, some required information in the query could appear close to the first sentence of the passage. If we include this previous sentence, then the comprehension of the passage is highly increased.

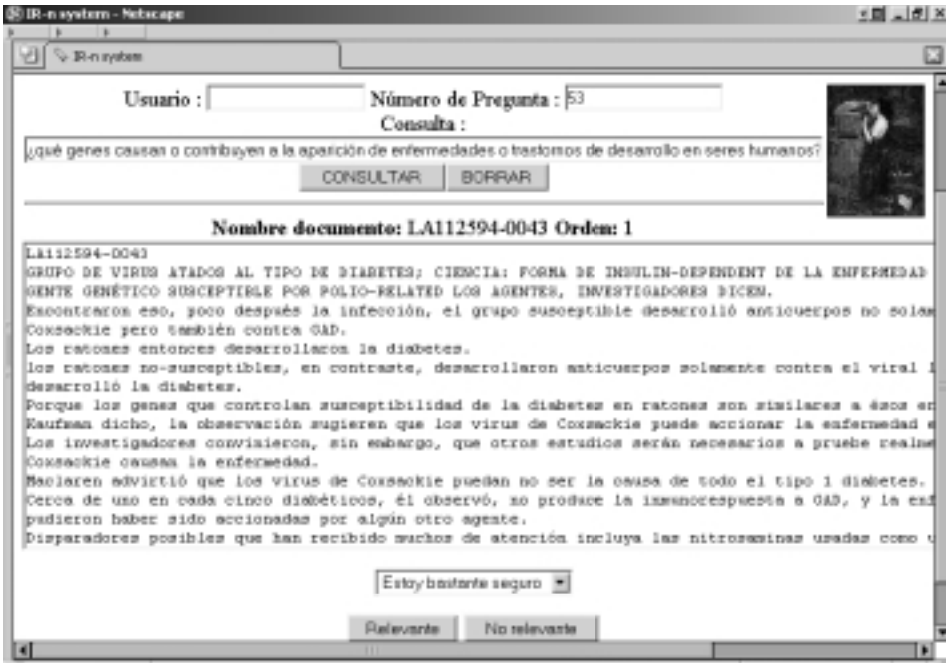

Fig. 1. HTML interface for presenting documents to the user.

3. The sentences are shown in different lines in the interface. This fact facilitates the comprehension of the passage, and it was quite easy to carry out since IR-n performs an indexation on sentences.

5 Results

The experiments were carried out by university students with a medium level of understanding of English (although it is not important since the passages are automatically translated into Spanish). The IR-n system results have been compared with those obtained with Z-Prise [3]. Only the first 25 most relevant documents have been used for each query, which could explain the low recall results. Finally, we would like to notice that only title and description of each query have been used. The obtained results have been presented in Table 1.

Table 1. Results comparison.

System	Average F_alpha
Z-Prise	0.2166
IR-n	0.3248

Given that only 25 relevant documents were retrieved for each topic, it is interesting to study the precision results. In Table 2, the precision results are presented for each topic, for both the Z-Prise and our system, IR-n. Firstly, it is remarkable the low results obtained for one topic, because any relevant document was retrieved by any system. Secondly, we should remark the high precision in two of the three remaining topics, which was obtained with just the most relevant passage.

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 relevance of documents, in which the users can only read the most relevant passages of these documents. The results have been quite good, because the users take short time to judge the relevance since they have to read short pieces of text. However, these short pieces of text contain the most relevant information about the information required, therefore the precision results have been high, even higher than those obtained by means of reading the entire document. Anyway, there are some points to notice, once the individual results and the opinions of the users have been analysed:

- Firstly, the users find great anxiety when they do not find the relevant document in the list of relevant passages. This has occurred in 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 been selected in other cases.
- Moreover, the users find the automatic translations into Spanish quite unreadable most of the times (more than it was expected).
- We think that the results have been influenced by presenting just the title and description of the queries, which have supposed 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 obtained results, since the users have 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.

As future works, firstly, we pretend 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 it was sometimes unreadable, we will try to present the results to the user in a more structured way. This task will be carried out by means of 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.

7 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*, pages 302–310, London, UK, July 1994. Springer Verlag.
2. CLEF2001. Workshop of cross-language evaluation forum (clef 2001). In *Workshop of Cross-Language Evaluation Forum (CLEF 2001)*, Lecture notes in Computer Science, Darmstadt, Germany, 2001. Springer-Verlag.
3. Darrin Dimmick. ZPrise information retrieval system. Avalaible 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)*, Lecture notes in Computer Science, pages 373–380, Mexico City, Mexico, 2002. Springer-Verlag.
6. José Luis Vicedo, Antonio Ferrández, and Fernando Llopis. University of Alicante at TREC-10. In *Tenth Text REtrieval Conference (Notebook)*, volume 500-250 of *NIST Special Publication*, Gaithersburg, USA, nov 2001. National Institute of Standards and Technology.