Passage Selection to Improve Question Answering

Fernando Llopis, Antonio Ferrández and José Luis Vicedo

Departamento de Lenguajes y Sistemas Informáticos
University of Alicante
Alicante, Spain
{llopis, antonio, vicedo}@dlsi.ua.es

Abstract. *Question Answering systems (QA) try to detect snippets of text in a collection of documents, which contain the response to a user's query. The complexity of QA systems reduces the applicability of these systems to smaller collections of documents. Therefore, QA systems usually employ different tools to reduce to text to work in, such as Information Retrieval (IR) systems. In this paper, we are proposing a Passage Retrieval tool in order to improve the precision and efficiency of a QA system as it was used in the last TREC-10 QA track. Here, we are evaluating this new tool against a standard IR system, and better results have been obtained.*

1 Introduction

Information Retrieval (IR) systems receive as input a user's query, and they have to return a set of documents sorted by their relevance to the query. There are different techniques to carry out the document extraction process, but most of them are based on pattern matching modules, where they calculate the number of times that a query term appear in each document, as well as they calculate the importance or weight of each term. Question Answering (QA) systems try to improve the output generated by IR systems by means of returning just snippets of text that are supposed to contain the response. QA systems uses Natural Language Processing as well as IR techniques in order to fully understand the text, and in this way, firstly they check if the document can contain the right answer and secondly they find the piece of text that contains it. Since these NLP techniques are computationally expensive, then QA systems have to reduce the amount of text to process. In this way, they usually work on the output of an IR system [10] or in the output of a Passage Retrieval (PR) system that selects the most relevant passages of each document [4]. Presently, several PR systems [2][5][8][9] have been proposed. These systems deal with fragments of text in order to determine if a document is relevant or not, and can use similar techniques as IR systems. IR systems are mainly based on three models: the cosine model [15], the pivoted cosine model[1] [17], and the probabilistic model called okapi [18]. Moreover, IR systems usually employ query expansion techniques that frequently improve their precision. These techniques can be based on thesaurus [21] or on the incorporation of the most frequent terms in the most relevant documents selected [7].

In [20], the QA system used in the TREC-9 (Text Retrieval Conference) QA track is presented. It worked on the first 50 documents returned by a standard IR system. In the TREC-10 [19], this QA system was improved by using the PR system called IR-n [11], and in spite of the increase in the difficulty of the questions, as it was expected, the obtained results were better than in TREC-9 (the mean reciprocal range rose 7 points).

In this paper, we are analysing the importance of the IR-n PR system for QA as it was used in TREC-10. The following section briefly presents the backgrounds in IR, PR and QA. Section 3 shows the architecture of IR-n. In Section 4, the evaluation is presented. Finally, we present the conclusions of this work.

2 Backgrounds in Question Answering and Passage Retrieval

2.1 Information Retrieval and Passage Retrieval

Let us suppose that the following question is posed to an IR system: *Who killed Lincoln?* The IR system would sort the documents by relevance to the query. Therefore, it would calculate the similarity between each document and the question by means of calculating the frequency of each query term in the document. This produces that bigger documents can be preferred. A possible alternative to these models of IR is the one that calculates the similarity in accordance with the relevance of the passages in the document. This new model of IR systems is called *Passage Retrieval* (PR), and its main advantage is that it is not affected by the length of each document as well as it obtains better precision. Moreover, PR systems allow QA systems to work with smaller pieces of text instead of whole documents. The improvement can be between 20 and 50% [2][9].

[1] It is a modification of the cosine model. It tries to reduce the problem of the preference for bigger documents.

Two classifications can be accomplished in PR. The first one is in accordance with the way of dividing the documents into passages. The second one is in accordance with the moment in which the passage segmentation is carried out. With reference to the first classification, it is generally agreed the one proposed in [2], where it distinguishes between models based on discourse, semantic models, and window models. The first one uses the structural properties of the documents, such as sentences or paragraphs (e.g. the one proposed in [13], [16]) in order to define the passages. The second one divides each document in semantic pieces, according to the different topics in the document [5]. The last one uses windows of a fixed size (usually a number of terms) to form the passages ([2], [8]).

It looks coherent that discourse-based models are more effective since they are using the structure of the document itself. However, the greater problem of them is that the results could depend on the writing style of the document author. 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. Nevertheless, discourse-based and semantic models have the main advantage that they return logic and coherent fragments of the document, which is quite important if these systems are used for other applications such as QA

According to the second PR classification, we can distinguish between those that previously segment the document into passages, and those that segment after the query is posed. The first one allows a quicker calculation, but the second one allows different segmentation models in accordance with the kind of query. The experiments presented in [9] show that the second one presents considerable advantages.

The passage extraction model that we are proposing, IR-n, allows us to benefit from the advantages from discourse-based models since logic information units of the text, such as sentences, form the passages. Moreover, another novel proposal in our PR system is the relevance measure, which unlike other discourse-based models, is not calculated from the number of passage terms, but the fixed number of passage sentences. This fact, allows a simpler calculation of this measure unlike other discourse-based or semantic models. Although we are using a fixed number of sentences for each passage, we consider that our proposal differs from the window models since our passages does not have a fixed size (i.e. a fixed number of words) because we are using sentences with a variable size. Furthermore, IR-n segments after the query is posed, which allows us to determine the number of sentences in the passage in accordance with the kind of query.

2.2 Question answering

Let us suppose that the previous question is also issued to a QA system: *Who killed Lincoln?* The QA system would search for the piece of text that contains the response to the question. Computational Linguistic community has shown a recent interest on QA, and it comes after developing Information Extraction systems, which have been evaluated in Message Understanding Conferences (MUC). Specifically, the interest was shown when in TREC-8, appears a new track on QA that tries to benefit from large-scale evaluation, that was previously carried out on IR systems, in previous TREC conferences.

If a QA system wants to successfully obtain a user's request, it needs to understand both texts and questions to a minimum level. That is to say, it has to carry on many of the typical steps on natural language analysis: lexical, syntactical and semantic. This analysis takes much more time than the only statistical analysis that is usually carried out in IR. Besides, as QA has to work with as much text as IR, and the user needs the answer in a limited interval of time, it is usual that an IR system processes the query and after, the QA system will continue with its output. In this way, the time of analysis is highly decreased.

Some of the best present QA systems are the following: [3][4][14][6]. After studying these systems, it seems agreeable the following general architecture, that is formed by four modules, where document retrieval module is accomplished by using IR technology:

- Question Analysis.
- Document Retrieval.
- Passage Selection.
- Answer Extraction.

3 IR-n overview

In this section, we describe the architecture of the proposed PR system, namely IR-n, focusing on its three main modules: the indexing, the document extraction and query expansion modules.

3.1 Indexing module

The main aim of this module is to generate the dictionaries that contain all the required information for the document-extraction module. It requires the following information for each term:

- The number of documents that contain the term.
- For each document:
  - The number of times that the term appears in the document.
  - The position of each term in the document: the number of sentence and position in the sentence.

Where we consider as terms, the stems produced by the Porter stemmer on those words that do not appear in a list of stop-words, list that is similar to those used in IR systems. For the query, the terms are also extracted in the same way, that is to say, their stems and positions in the query for each query word that does not appear in the list of stop-words.

3.2 Document extraction module

This module extracts the documents according to its similarity with the user's query. The scheme in this process is the following:

1. Query terms are sorted according to the number of documents in which they appear, where the terms that appear in fewer documents are processed firstly.
2. The documents that contain some query term are extracted.
3. The following similarity measure is calculated for each passage *p* with the query *q*:

Similarity_measure(p, q) = $\sum_{t \in p \wedge q} W_{p,t} * W_{q,t}$

Where:

$W_{p,t} = \log_e(f_{p,t} + 1)$.

$f_{p,t}$ is the number of times that the term *t* appears in the passage *p*.

$W_{q,t} = \log_e(f_{q,t} + 1) * \text{idf}$.

$f_{q,t}$ is the number of times that the term *t* appears in the query *q*.

$idf = \log_e(N / f_t + 1)$.

N is the number of documents in the collection.

f_t is the number of documents that contain the term *t*.

4. Each document is assigned the highest similarity measure from its passages.
5. The documents are sorted by their similarity measure.
6. The documents are presented according to their similarity measure.

As it is noticed, the similarity measure is comparable to cosine measure presented in [15]. The only difference is that the size of each passage (the number of terms) is not used to normalise the results. This difference makes the calculation simpler than other discourse-based PR systems or IR systems, since the normalization is accomplished according to a fixed number of sentences per passage. Another important detail to notice is that we are using *N* as the number of documents in the collection, instead of the number of passages. That is because in [9] it is not considered relevant for the final results.

The optimum number of sentences to consider per passage is experimentally obtained. It can depend on the genre of the documents, or even on the type of the query as it is suggested in [8].

As it is commented, the proposed PR system can be classified into discourse-based models since it is using variable-sized passages that are based on a fixed number of sentences (but different number of terms per passage). The passages overlap each other, that is to say, let us suppose that the size of the passage is *N* sentences, then the first passage will be formed by the sentences from 1 to N, the second one from 2 to N+1, and so on. We have decided to overlap just one sentence based on the experiments accomplished in [12] in order to calculate the optimum number of overlapping sentences in each passage, where only one overlapping sentence obtained the best results.

3.3 Query expansion modules

In IR-n several query expansion techniques were tested. Firstly, synonyms obtained from WordNet were used to expand the query, but worst results were obtained than without query expansion in [11]. The issue and problems of query expansion is also studied in [3]. Presently, IR-n is using the blind relevance feedback model in order to expand the query as it is described in [1], in which it introduces the 15 most frequent terms in the five first most relevant passages of 10 sentences.

4 Evaluation

This section presents the experiment proposed for evaluating our approach and the results obtained. The experiment has been run on the TREC-9 QA Track question set and document collections.

4.1 Data collection

TREC-9 question test set is made up by 682 questions with answers included in the document collection. The document set consists of 978,952 documents from the TIPSTER and TREC following collections: AP Newswire, Wall Street Journal, San Jose Mercury News, Financial Times, Los Angeles Times, Foreign Broadcast Information Service.

4.2 Experiment

In order to evaluate our proposal we decided to compare the quality of the information retrieved by our system with the ranked list retrieved by the ATT IR system. Firstly, ATT IR system was used for retrieving the first 1000 relevant documents for each question. In order to measure the relevance of these documents, Table 1 shows the results of the following experiment. This Table shows the number of questions whose answer can be found in the first *n* documents returned by the ATT IR system. These results are divided into two columns: one for only 100 questions, and other one for all the 682 questions evaluated in TREC-9 QA track. This is for determine a training set of 100 questions, in which several experiments were carried out to fine-tune our system.

n	100 questions	682 questions
5	62%	442 (64.90 %)
10	69%	479 (70.33 %)
20	77%	517 (75.91%)
30	82%	539 (79.14%)
50	83%	570 (83.70 %)
100	87%	595 (87.37 %)
200	89%	613 (90.01%)
500	92%	631 (92.65%)

Table 1. Questions rightly answered in the *n* fist documents returned by ATT IR system.

One of these training experiments consists of working on the output of the ATT system, in order to re-sort its output using IR-n. Another experiment consists of using IR-n as the main IR system, i.e. indexing the whole collections by means of IR-n. In each experiment, a different number of sentences in each passage was tested: 5, 10, 15 and 20 sentences. The relevance of each returned document was measured by means of the tool provided by TREC organization that allows us to determine if a passage contains the right answer. The two experiments are summed up in Figure 1.

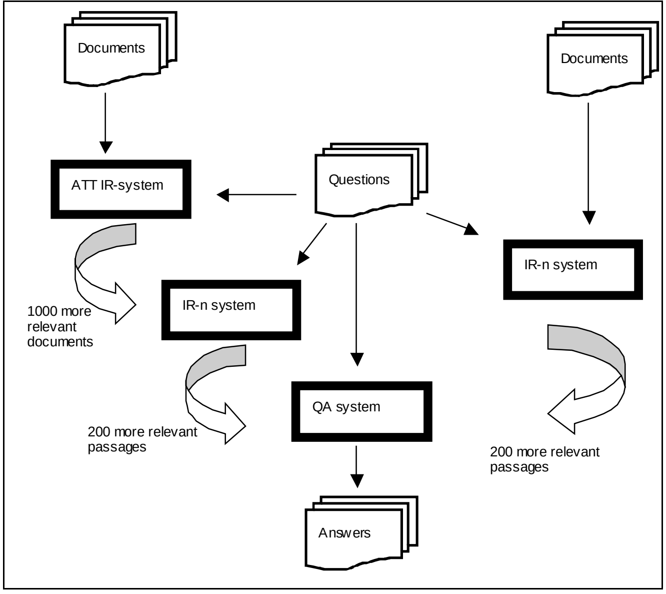
Documents
Documents
ATT IR-system
Questions
IR-n system
1000 more relevant documents
IR-n system
QA system
200 more relevant passages
200 more relevant passages
Answers

Figure 1. Description of experiments

4.3 Results obtained

Table 2 shows the obtained results for passages of 5, 10, 15 and 20 sentences using different systems. The first system (IR-n Ref) is IR-n when it works on the 1000 documents returned by ATT system. The following two systems work overall collections, where the second one (IR-n QE) uses the blind relevance feedback model to expand the queries, whereas the first one (IR-n) does not. Finally, the results obtained by ATT system (previously shown in Table 1) are presented.

Answer Included	At 5 docs	At 10 docs	At 20 docs	At 30 docs	At 50 docs	At 100 docs	At 200 docs
IR-n Ref.							
5 Sentences	57	66	78	83	85	88	93
10 Sentences	63	76	80	89	93	96	97
15 Sentences	*70*	*78*	*83*	*89*	*94*	*95*	*96*
20 Sentences	74	83	87	91	93	96	97
IR-n							
5 Sentences	55	63	75	80	84	89	90
10 Sentences	60	73	78	87	92	95	97
15 Sentences	*70*	*76*	*82*	*87*	*93*	*95*	*95*
20 Sentences	72	80	86	90	92	96	96
IR-n . QE							
5 Sentences	52	60	70	72	74	78	88
10 Sentences	59	68	77	79	82	84	91
15 Sentences	*62*	*74*	*82*	*82*	*83*	*86*	*93*
20 Sentences	65	75	83	86	87	89	94
ATT system							
	62	69	77	82	83	87	89

Table 2. Number of questions rightly answered (training set of 100 questions).

From these results, the following conclusions can be extracted. Firstly, IR-n Ref obtains similar results than IR-n, i.e. it is not improved when it works on the whole collections. Moreover, expansion query techniques have not obtained the expected improvement, although IR-n always improves the sorting provided by ATT-system. Finally, it can be observed that the best results are obtained with passages of 15 and 20 sentences, and that the percentage obtained with 200 documents is quite acceptable. Furthermore, if we want to improve this percentage, the number of document is highly increased.

After the training stage, the whole set of questions are resolved, where IR-n Ref (with passages of 15 and 20 sentences) is compared with ATT system, whose results are shown in Table 3. In Figure 2. is compared ATT-system with IR-n system using passages of 20 sentences. As it can be observed, our system reaches a maximum improvement of 12 points with reference to ATT system when only 20 documents are returned.

Answer included	ATT system	15 Sent.	20 Sent.
At 5 docs	442 (64.90%)	*488 (71.65%)*	508 (74.59%)
At 10 docs	479 (70.33%)	*549 (80.61%)*	561 (82.73%)
At 20 docs	517 (75.91%)	*584 (85.75%)*	595 (87.37%)
At 30 docs	539 (79.14%)	*600 (88.10%)*	612 (89.96%)
At 50 docs	570 (83.70%)	*623 (91.48%)*	624 (91.62%)
At 100 docs	595 (87.37%)	*640 (93.97%)*	644 (94.56%)
At 200 docs	613 (90.01%)	*648 (95.15%)*	654 (96.03%)

Table 3. ATT-system versus IR-n system.

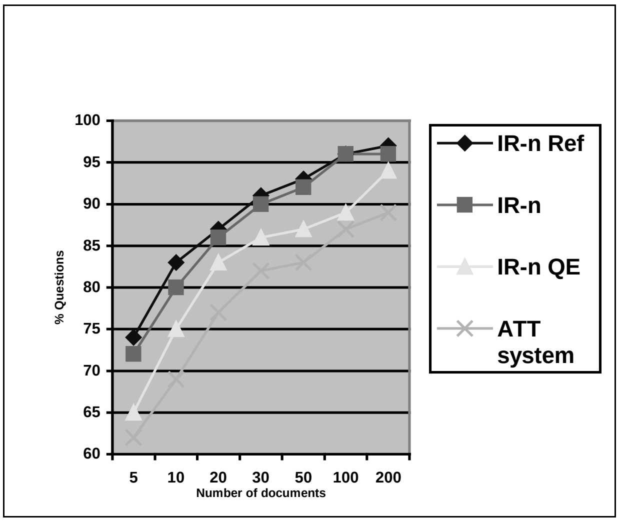

Figure 2. Comparative of ATT-system and experiments with IR-n (Passages of 20 sentences)

5 Conclusions and future works

In this paper, we have analysed the improvement obtained by our Passage Retrieval (PR) system, called IR-n, with reference to a standard IR system, ATT IR system. This improvement has been measured on the TREC-9 QA track questions. The improvement consists on a better precision, and in reducing the amount of text that the QA system has to work with. In the presented experiments, IR-n has shown similar results when it works on the output of ATT system, than when it works on the whole collections. Moreover, it has proved to work better with passages of 15 and 20 sentences, with a maximum improvement of 12 points with reference to ATT system when only 20 documents are returned. Furthermore, it is important to remark that query expansion techniques have not obtained better results. Finally, just mention that IR-n also allows us to work on smaller pieces of text, i.e. the QA system can only work on the passages instead of the whole document. In this way, an improvement of 7 points in the mean reciprocal range was obtained in TREC-10.

As future works, we intend to determine the optimum size of passages in accordance with the kind of question, in order to improve the precision of the system. Moreover, we are trying to improve the relationship between IR-n and the following QA system, in order to detect the maximum number of passages to extract for each query.

References

[1] Bertoldi, N and Federico, M. *ITC-irst at CLEF-2001 , Working Notes for the Clef 2001* Darmstdt, Germany , pp 41-44

[2] Callan, J. *Passage-Level Evidence in Document Retrieval*. In Proceedings of the 17 th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland 1994, pp. 302-310.

[3] Clarke, C.; Cormack, g, Kisman, D and Lynam, T. *Question Answering by Passage Selection(Multitext Experiments for TREC-9*) Proceedings of the Tenth Text REtrieval Conference, TREC-9. Gaithersburg , USA 2000, pp 673-683

[4] Harabagiu, S.; Moldovan, D.; Pasca, M.; Mihalcea, R.; Surdeanu, M.; Bunescu, R.; Gîrju, R.; Rus, V. and Morarescu, *P. FALCON: Boosting Knowledge for Answer Engines*. In Nineth Text REtrieval Conference, Gaithersburg *USA 2000.pp 479-*

[5] Hearst, M. and Plaunt, C. *Subtopic structuring for full-length document access.* Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA USA 1993 , pp 59-68

[6] *Ittycheriah, A.; Franz, M.; Zu, W. and Ratnaparkhi, A. IBM's Statistical Question Answering System.* In Nineth Text REtrieval Conference, Gaithersburg *USA 2000.*, pp 231-236

[7] J. Xu and W. Croft. *Query expansion using local and global document analysis.* In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 1996 pp 4—11, 18—22.

[8] Kaskiel, M. and Zobel, J. *Passage Retrieval Revisited* SIGIR '97: Proceedings of the 20th Annual International ACM Philadelphia, PA 1997, USA, pp 27-31

[9] KaszKiel, M. and Zobel, J. *Effective Ranking with Arbitrary Passages.* Journal of the American Society for Information Science, Vol 52, No. 4, February 2001, pp 344-364.

[10] Litkowski, k, Syntactic Clues and Lexical Resources in Question-Answering *In Nineth Text REtrieval Conference,* Gaithersburg *USA 2000* pp177-188

[11] Llopis, F. and Vicedo, J. *Ir-n system, a passage retrieval system at CLEF 2001* Working Notes for the Clef 2001 Darmstdt, Germany 2001, pp 115-120 . To appear in Lecture Notes in Computer Science

[12] Llopis, F.; Ferrández, and Vicedo, J. *Text Segmentation for efficient Information Retrieval* Third International Conference on Intelligent Text Processing and Computational Linguistics. Mexico 2002 To appear in Lecture Notes in Computer Science

[13] Namba, I *Fujitsu Laboratories TREC9 Report.* Proceedings of the Nineth Text REtrieval Conference, TREC-9. Gaithersburg,USA.2000, pp 203-208

[14] Prager, J.; Brown, E.; Radev, D. and Czuba, K. *One Search Engine or Two for QuestionAnswering.* In *Nineth Text REtrieval Conference*, Gaithersburg,USA. 2000.

[15] Salton G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer,* Addison Wesley Publishing, New York. 1989

[16] Salton, G.; Allan, J. Buckley *Approaches to passage retrieval in full text information systems.* In R Korfhage, E Rasmussen & P Willet (Eds.) Prodeedings of the 16 th annual international ACM-SIGIR conference on research and development in information retrieval. Pittsburgh PA USA , pp 49-58

[17] Singhal, A.; Buckley, C. and Mitra, M. *Pivoted document length normalization.* Proceedings of the 19th annual international ACM-SIGIR conference on research and development in information retrieval, 1996**.**

[18] Venner, G. and Walker, S. *Okapi '84: `Best match' system.* Microcomputer networking in libraries II. Vine, 48,1983, pp 22-26.

[19] Vicedo, J.; Ferrandez, A and Llopis, F. *University of Alicante al TREC-10. In Tenth Text REtrieval Conference,* Gaithersburg,USA. *2001*

[20] Vicedo, J.; Ferrandez, A; *A semantic approach to Question Answering systems. In Nineth Text REtrieval Conference, 2000 pp 440-444.*

[21] Y. Jing and W. B. Croft. *An* ***association thesaurus*** *for* ***information retrieval***. In RIAO 94 Conference Proceedings, , New York, 1994. pp 146--160