Passage Selection to Improve Question Answering

Fernando LLopis
Departamento de Lenguajes y Sistemas Informáticos
Alicante (Spain) 03800
llopis@dlsi.ua.es

José Luis Vicedo
Departamento de Lenguajes y Sistemas Informáticos
Alicante (Spain) 03800
vicedo@dlsi.ua.es

Antonio Ferrández
Departamento de Lenguajes y Sistemas Informáticos
Alicante (Spain) 03800
antonio@dlsi.ua.es

Abstract

Open-Domain Question Answering systems (QA) performs the task of detecting text fragments in a collection of documents that contain the response to user's queries. These systems use high complexity tools that reduce its applicability to the treatment of small amounts of text. Consequently, when working on large document collections, QA systems apply Information Retrieval (IR) techniques to reduce drastically text collections to a tractable quantity of relevant text. In this paper, we propose a novel Passage Retrieval (PR) model that performs this task with better performance for QA purposes than current best IR systems

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 that depend on the number of times that a query term appear in each document, as well as the importance or discrimination value of each term in the document collection. Question Answering (QA) systems try to improve the output generated by IR systems by means of returning just small pieces of text that are supposed to contain the response. Usually, QA systems combine IR and Natural Language Processing (NLP) techniques to perform their task. This combination allows text understanding until a minimum level that permits a precise answer detection and extraction. Nevertheless, since NLP techniques are computationally expensive, QA systems need to reduce the amount of text where these techniques have to be applied. In this way, they usually work on the output of IR systems [10] that select the most relevant documents to the query by supposing that they will contain the answer required. Most applied IR systems are mainly based on three models: the cosine model [15], the pivoted cosine model[1] [17], and the probabilistic model (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 top M relevant documents [7].

Currently, several Passage Retrieval (PR) systems have also been proposed for this task [2][5][8][9]. PR systems deal with fragments of text in order to determine the relevance of a document to a query, as well as to detect document extracts that are likely to contain the expected answer (instead of full documents). Although PR systems apply IR-based techniques to perform their work, they have revealed to be more effective than IR systems for QA tasks.

In this paper, we are analysing the importance of the IR-n PR system for QA n [11] as it was used in last TREC-10 Conference [19]. The following section briefly presents the backgrounds in IR, PR and QA. Section 3 shows the architecture of IR-n. Section 4 presents the evaluation accomplished and finally, section 5 details conclusions and work in progress.

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

2 Backgrounds in Question Answering and Passage Retrieval

2.1 Information Retrieval and Passage Retrieval

Given a question, an IR system sorts the documents by its relevance to the query. It computes the similarity between each document and the question by taking into account the frequency of each query term in the document. This fact usually produces that bigger documents are preferred. A possible alternative to IR models is based on obtaining the similarity in accordance with the relevance of the passages contained in the document. This new approach, called *Passage Retrieval* (PR), has several advantages. When used for document retrieval, as the relevance of a document will depend on the relevance of the passages it contains, this measure will not be affected by the length of the full document. Moreover, these techniques allow to detect high relevant information embedded in a long document obtaining, this way, better performance than IR approaches [2][9]. On the other hand, when applied for QA tasks, PR systems allow reducing the amount of text to be processed with costly NLP tools by returning passages instead of whole 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 one, PR community generally agrees with the classification proposed in [2], where the author distinguishes between discourse models, semantic models, and window models. The first one uses the structural properties of the documents, such as sentences or paragraphs [13][16] in order to define the passages. The second one divides each document into 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 determine passage boundaries [2] [8].

At first glance, we could think that discourse-based models would be the most effective, in retrieval terms, since they use the structure of the document itself. However, this model greatest problem relies on detecting passage boundaries since it depends on the writing style of the author of each document. On the other hand, window models have as 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 full information units of the document, which is quite important if these units are used as input by other applications such as QA.

According to the second classification, we can distinguish between approaches that segment documents into passages for indexing purposes, and those that perform segmentation after the query is posed. The first one allows a quicker calculation; nevertheless, the second one allows different segmentation models in accordance with the kind of query.

The passage extraction model that we propose (IR-n) allows us to benefit from the advantages of discourse-based models since self-contained information units of text, such as sentences, are used for building passages. Moreover, another novel proposal in our PR system is the relevance measure which, unlike other discourse-based models, is not based on the number of passage terms, but on a fixed number of passage sentences. This fact allows a simpler calculation of this measure unlike other discourse-based or semantic models. Although each passage is made up by a fixed number of sentences, we consider that our proposal differs from the window models since our passages do not have a fixed size (i.e. a fixed number of words) since we use sentences with a variable size. Furthermore, IR-n document segmentation into passages is accomplished after the query is posed, which allows us to determine the number of sentences to be considered in accordance with the kind of the query.

2.2 Question Answering

Open domain QA systems are defined as tools capable of extracting the answer to user queries directly from unrestricted domain documents. Or at least, systems that can extract text snippets from texts, from whose content it are possible to infer the answer to a specific question. In both cases, these systems try to reduce the amount of time users spend to locate a concrete information.

Interest in QA systems is quite recent. We had little information about this kind of systems until the “First Question Answering Track” was held in TREC-8 Conference. This track 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. From a linguistic perspective, “understanding” means to carry out many of the typical steps on natural language analysis: lexical, syntactic and semantic. This analysis takes much more time than the statistical analysis that is usually carried out in IR. Besides, as QA systems have to manage with as much text as done for IR tasks, and the user needs the answer in a limited interval of time, it is nearly mandatory that first, an IR system processes the query and second, the QA process continues with its output. In this way, the time of analysis is highly decreased.

The analysis of current best systems [3] [4] [14] [6] allows identifying main QA sub-components where document retrieval 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: indexing, passage retrieval and query expansion.

3.1 Indexing module

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

- The number of documents that contain this term.
- For each document:
  - The number of times this term appears in the document.
  - The position of each term in the document represented as the number of sentence it appears in.

As term, we consider the stem produced by the Porter stemmer on those words that do not appear in a list of stop-words, list that is similar to those generally used for IR. On the other hand, query terms are also extracted in the same way, that is to say, we only consider the stems of query words that do not appear in the stop-words list.

3.2 Passage retrieval module

This module extracts the passages 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 they appear in. Terms that appear in fewer documents are processed firstly.
2. The documents that contain any query term are selected.
3. The following similarity measure is calculated for each passage p (contained in the selected documents) with the query q:

$$\text{Similarity\_measure}(p, q) = \sum_{t \in p \wedge q} \mathrm{W}_{\mathrm{p,t}} \cdot \mathrm{W}_{\mathrm{q,t}}$$

$$W_{p,t} = \log_e(f_{p,t} + 1).$$
$$W_{q,\,t} = \log_e(f_{q,t} + 1) \cdot idf$$
$$idf\ = \log_e(N / f_t + 1)$$

Where $f_{p,t}$ is the number of times that the term t appears in the passage p. $f_{q,t}$ represents the number of times that the term t appears in the query q. N is the number of documents in the collection and f_t is refers to the number of documents that contain the term t.

4. Only the most relevant passage of each document is selected for retrieval.
5. The selected passages are sorted by their similarity measure.
6. Passages are associated with the document they pertain and they are presented in a ranked list form.

As we can notice, the similarity measure is similar to the 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 proposal performs normalization according to the fixed number of sentences per passage. This difference makes the calculation simpler than other discourse-based PR or IR systems. Another important detail to remark is that we are using *N* as the number of documents in the collection, instead of the number of passages according to the considerations presented in [9].

As it has been commented, our PR system uses variable-sized passages that are based on a fixed number of sentences (with different number of terms per passage). The passages overlap each other, that is to say, if a passage contains *N* sentences, 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 decided to overlap just one sentence according to the experiments and results presented in [12]. This work studied the optimum number of overlapping sentences in each passage for retrieval purposes concluding, that best results were obtained when only one overlapping sentence was used. Regarding to the optimum number (*N*) of sentences per passage considered in this paper, it will be experimentally obtained.

4 Evaluation

This section presents the experiments developed for training and evaluating our approach. The experiments have 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 Training

Training experiments had two objectives. They were designed (1) to calculate the optimum number of sentences (*N*) that define passage length and (2) to test two different possible ways of applying our method.

First training experiment consists of working on the output of one of the current best performing IR systems (the ATT system). This experiment re-sorts its output (the first 1,000 ranked documents) by using IR-n. Second experiment consists of using our proposal as the main IR system, that is, indexing the whole collections by means of IR-n. For each experiment, a different number of sentences per passage were 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.

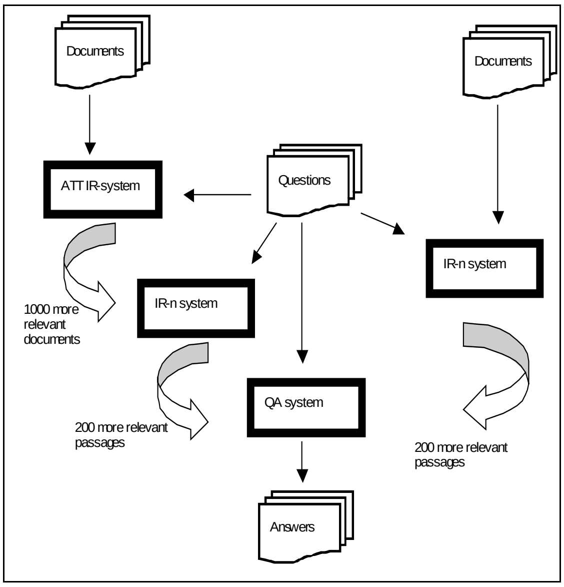
Documents
Documents
ATT IR-system
Questions
IR-n system
IR-n system
1000 more relevant documents
QA system
200 more relevant passages
200 more relevant passages
Answers

Figure 1. Training Experiments

These experiments were performed using only the first 100 questions included in the data collection. Table 1 shows training results for passages of 5, 10, 15 and 20 sentences using both approaches. This results measure the number of questions whose correct answer was included into the top *n* retrieved passages (or documents) for the training question set. The first experiment (*IR-n Ref*) uses IR-n on the 1,000 documents returned by ATT system while the second one (*IR-n*) applies passage retrieval overall collections.

As we can see, *IR-n Ref* and *IR-n* test obtain similar results although using our approach to re-rank the output of a good IR system presents a slight better performance than applying *IR-n* overall document collection. Regarding to the number of sentences to be taken into account to define passage length, we can observe that best results are obtained with passages of 20 sentences. In this case, both tests improve significantly the performance of ATT-system. It ranges from 12 (*IR-n Ref*) and 10 (*IR-n*) points on a passage length of 20 sentences (for only the first 5 documents retrieved) to 8 and 7 points when the first 200 documents are taken into account respectively.

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 Sent	57	66	78	83	85	88	93
10 Sent	63	76	80	89	93	96	97
15 Sent	*70*	*78*	*83*	*89*	*94*	*95*	*96*
20 Sent	74	83	87	91	93	96	97
IR-n							
5 Sent	55	63	75	80	84	89	90
10 Sent	60	73	78	87	92	95	97
15 Sent	*70*	*76*	*82*	*87*	*93*	*95*	*95*
20 Sent	72	80	86	90	92	96	96
ATT system							
	62	69	77	82	83	87	89

Table 1. Number of questions rightly answered (training set of 100 questions).

4.3 Experiment

In order to evaluate our proposal we decided to compare the quality of the information retrieved by our approaches with the ranked list retrieved by the ATT IR system. For this evaluation, the 682 questions included in the data collection were processed and the number *N* of sentences per passage was set to 20. Table 2 shows the results of this evaluation experiment. This table shows the percentage of questions whose answer can be found into the first *n* documents returned by the ATT IR system and the best *n* passages returned by *IR-n Ref* and *IR-n* respectively. These results are also presented in Figure 2

These data confirm training results. In this case, both approaches perform better than ATT system and improvements range form 6 to 12 points for 20 sentences passage length.

Answer Included	ATT system	IR-n Ref	IR-n
At 5 docs	64.90%	74.59%	72.21%
At 10 docs	70.33%	82.73%	80.37%
At 20 docs	75.91%	87.37%	86.35%
At 30 docs	79.14%	89.96%	89.31%
At 50 docs	83.70%	91.62%	91.52%
At 100 docs	87.37%	94.56%	95.55%
At 200 docs	90.01%	96.03%	95.92%

Table 2. ATT-system versus IR-n systems.

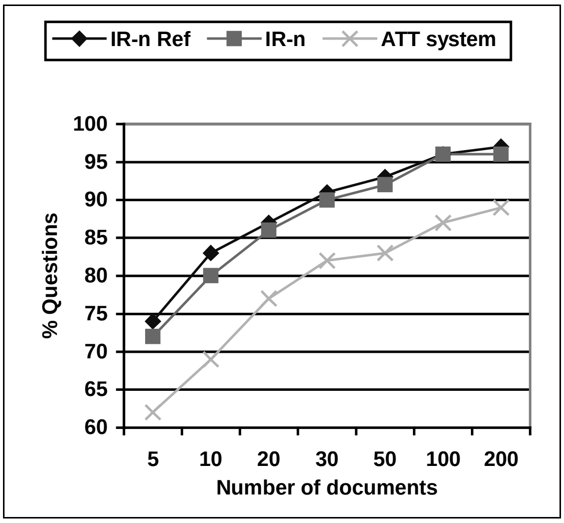

Figure 2. Comparative of ATT-system and experiments with IR-n (Passages of 20 sentences)

5 Conclusions and future works

In this paper, we have analysed the improvement obtained by our passage retrieval system, called IR-n, with reference to a high-performance IR system (ATT) regarding to is application for QA tasks. This improvement has been evaluated on the TREC-9 QA track data set. The achieved improvements are twofold: First, our approach obtains a better precision by retrieving more passages that contain the answer to users' queries than ATT system does. Second, since our approach returns passages (instead of documents), it significantly reduces the amount of text to be processed with costly techniques by the QA system. The related experiments show that the optimal passage length for this task is 20 when passages are made up by a fixed number of sentences. Moreover, we have tested two different ways of applying our model. As we have seen, IR-n presents similar results when it works on the output of an IR system, than when it works on the whole collections. Nevertheless, in both cases, benefits range from 6 to 12 points with reference to ATT system depending on the number of first documents or passages retrieved to be processed for QA tasks.

As future work, in order to improve our system precision, we intend to obtain the optimum size of passages in accordance with the kind of question. Besides, we need to investigate the effects of query expansion techniques on IR-n system. Furthermore, we are also trying to improve the relationship between IR-n and the following QA system, in order to detect the minimum number of passages to extract for each query without affecting QA performance.

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, 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- 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