Language Independent Passage Retrieval for Question Answering

José Manuel Gómez-Soriano[1], Manuel Montes-y-Gómez[2], Emilio Sanchis-Arnal[1], Luis Villaseñor-Pineda[2], Paolo Rosso[1]

[1] Polytechnic University of Valencia, Spain.
{jogomez,esanchis,prosso}@dsic.upv.es
[2] National Institute of Astrophysics, Optics and Electronics, Mexico.
{mmontesg, villasen}@inaoep.mx

Abstract. Passage Retrieval (PR) is typically used as the first step in current Question Answering (QA) systems. Most methods are based on the vector space model allowing the finding of relevant passages for general user needs, but failing on selecting pertinent passages for specific user questions. This paper describes a simple PR method specially suited for the QA task. This method considers the structure of the question, favoring the passages that contain the longer *n*-gram structures from the question. Experimental results of this method on Spanish, French and Italian show that this approach can be useful for multilingual question answering systems.

1 Introduction

The volume of online available information is growing every day. Complex information retrieval (IR) methods are required to achieve the needed information. QA systems are IR applications whose aim is to obtain specific answers for natural language user questions.

Passage Retrieval (PR) is typically used as the first step in current QA systems [1]. Most of these systems apply PR methods based on the classical IR vector space model [2, 3, 4, 5], allowing the finding of relevant passages for general user needs, but failing on selecting pertinent passages for specific user questions. These methods use the question keywords in order to find relevant passages. For instance, for the question "*Who is the president of Mexico?*", they return a set of passages containing the words *president* and *Mexico*, but not necessarily a passage with the expected answer.

In [6, 7] it is shown that standard IR engines (such as MG and Okapi) often fail to find the answer in the documents (or passages) when presented with natural language questions. On the contrary, PR approaches based on Natural Language Processing (NLP) produce results that are more accurate [9, 10, 11, 12]. However, these approaches are difficult to adapt to several languages or to multilingual tasks.

Another common strategy for QA is to search the obviousness of the answer in the Web [13, 14, 15]. The idea is to run the user question into a Web search engine (usually Google) with the expectation to get a passage –snippet– containing the same expression of the question or a similar one. The methods using this approach suppose that due to high redundancy of the Web, the answer is written in several different

ways including the same form of the question. To increase the possibility to find relevant passages they make reformulations of the question, i.e., they move or delete terms to search other structures with the same question terms. For instance, they produce the reformulation "*the president of Mexico is*" for the question "*Who is the president of Mexico?*". Thanks to the redundancy, it is possible to find a passage with the structure "*the president of Mexico is* ***Vicente Fox***"'.

[14] makes the reformulations carrying out a Part Of Speech analysis of the question and moving or deleting terms of specific morph-syntactic categories. Whereas [13] makes the reformulations without doing any linguistic analysis, but just considering certain assumptions about the function of the words, such as the first or second question term is a verb or an auxiliary verb.

The problem of these methods is that not all possible reformulations of the question are considered. With these methods, it would be very costly to realize all possible reformulations, since the search engine must search for every reformulation. Our QA-oriented PR system makes a better use of the document collection redundancy bearing in mind all possible reformulations of the question efficiently running the search engine with just one question. Later the system searches for all word sequences of the question in the returned passages and weights every passage according to the similarity with the question. The passages with the more and the greater question structures will obtain better similarity values.

Moreover, given that our PR method does not involve any knowledge about the lexicon and the syntax of the specified language, it can be easily adapted to several different languages. It is simply based on the "superficial" matching between the question and the passages. As a result, it would work very well in any language with few differences between the question and the answer passages. In other words, it would be adequate for moderately inflected languages like English, Spanish, Italian and French, but not for agglutinative languages such as German, Japanese, and Nahuatl.

This paper presents the basis of our PR system and demonstrates it language independence condition with some experiments on three different languages. It is organized as follows. The section 2 describes the general architecture of the system and the equations. The section 3 discusses the experimental results of the method on Spanish, French and Italian. Finally, the section 4 presents our preliminary conclusions.

2 Passage Retrieval System

2.1 Architecture

The architecture of our PR system is shown in the figure 1.

Given a user question, it is firstly transferred to the *Search Engine* module. The Search Engine finds the passages with the relevant terms (non-stopwords), using a classical IR technique based on the vector space model. This module returns all passages that contain some relevant terms, but since the *n*-gram extraction is computationally expensive, it is necessary to reduce the number of passages for the *N-grams Extraction* module. Therefore, we only take, typically, the first 1000 passages (pre-

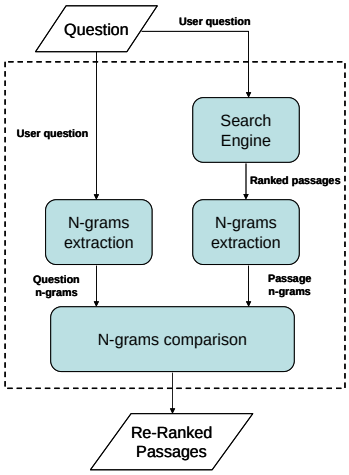

Figure 1. Diagram of the PR system

vious experiments have demonstrated that this is an appropriated number since it covers, in most of the cases, the whole set of relevant passages).

Once the passages are obtain by the *Search Engine* module, the sets of unigrams, bigrams,..., *n*-grams are extracted from the passages and from the user question by means of the *N-grams Extraction* modules. In both cases, *n* is the number of question terms.

Then, the *N-grams Comparison* module measures the similarity between the *n*-gram sets of the passages and the user question in order to obtain the new weights for the passages. The weight of a passage is related to the lager *n*-gram structure of the question that can be found in the passage itself. The larger the *n*-gram structure, the greater the weight of the passage.

Finally, the passages with the new weights are returned to the user.

2.2 Passage Ranking

The similarity between a passage *d* and a question *q* is defined by (1).

$$sim(d,q) = \frac{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x, D_j)}{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x, Q_j)} \quad (1)$$

Where $sim(d, q)$ is a function which measures the similarity of the set of *n*-grams of the question *q* with the set of *n*-grams of the passage *d*. Q_j is the set of *j*-grams that are generated from the question *q* and D_j is the set of *j*-grams of the passage *d* to compare with.

That is, Q_1 will contain the question unigrams whereas D_1 will contain the passage unigrams, Q_2 and D_2 will contain the question and passage bigrams respectively, and so on until Q_n and D_n.

The result of (1) is equal to 1 if the longest *n*-gram of the question is in the set of passage *n*-grams.

The function $h(x, D_j)$ measures the relevance of the *j*-gram *x* with respect to the set of passage *j*-grams, whereas the function $h(x, Q_j)$ is a factor of normalization. The function *h* assigns a weight to every question *n*-gram as defined in (2).

$$h(x, D_j) = \begin{cases} \sum_{k=1}^{n} w_k & \text{if } x \in D_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where $w_1, w_2, ..., w_{|x|}$ are the associated weights of the terms of the *j*-gram *x*.

These weights give an incentive to those terms that appear rarely in the document collection. Moreover, the weights should also discriminate the relevant terms against those (e.g. stopwords) which often occur in the document collection.

The weight of a term is calculated by (3):

$$w_k = 1 - \frac{\log(n_k)}{1 + \log(N)} \quad (3)$$

Where n_k is the number of passages in which appears the term associated to the weight w_k and *N* is the total number of passages in the collection. We assume that the stopwords occur in every passage (i.e., n_k takes the value of *N*).

For instance, if the term appears once in the passage collection, its weight will be equal to 1 (the maximum weight), whereas if the term is a stopword, then its weight will be the lowest.

2.3 Example

Assume that the user question is "*Who is the president of Mexico?*" and that we obtain two passages with the following texts: "*Vicente Fox is the president of Mexico…*" (p_1) and "*The president of Spain visited Mexico in last February…*" (p_2).

If we split the original question into five sets of *n*-grams (5 is the number of question terms without the question word *Who*) we obtain the following sets:

5-gram: "is the President of Mexico".

4-gram: "is the President of", "the President of Mexico".

3-gram: "is the President", "the President of", "President of Mexico".

2-gram: "is the", "the President", "President of", "of Mexico".

1-gram: "is", "the", "President", "of", "Mexico".

Next, we obtain the five sets of *n*-grams from the two passages. The passage p_1 contains all the *n*-grams of the question (the one 5-gram, the two 4-grams, the three 3-grams, the four 2-grams and the five 1-grams of the question). If we calculate the similarity of the question with this passage, we obtain a similarity of 1.

The sets of *n*-grams of the passage p_2 contain only the "*the President of*" 3-gram, the "*the President*" and "*President of*" 2-grams and the following 1-grams: "*the*", "*President*", "*of*" and "*Mexico*". If we calculate (1) for this passage, we obtain a

similarity of 0.29, a lower value than for p_1 because the second passage is very different with respect to the original question, although it contains all the relevant terms of the question.

3 Experimental Results

This section presents some experimental results on three different languages: Spanish, Italian and French. The experiments were carried out using the CLEF-2004[1] data set. This data set contains a corpus of news documents for each language as well as a list of several questions and their corresponding answers. Table 1 shows some numbers from the document corpora.

Table 1. Corpora statistics

	# documents	# sentences	# words
Spanish	454,045	5,636,945	151,533,838
Italian	157,588	2,282,904	49,343,596
French	129,806	2,069,012	45,057,929

For the experiments detailed in this section, we considered only the subset of factual questions (the questions having a named entity, date or quantity for answer) stated on the Multi-Eight CLEF04 question set having an answer in the Spanish, Italian or French document corpora.

For the evaluation we used a metric know as coverage (for more details see [7]). Let Q be the question set, D the passage collection, $A_{D,q}$ the subset of D containing correct answers to $q \in Q$, and $R_{D,q,n}$ be the top n ranked documents in D retrieved by the search engine given a question q. The coverage of the search engine for a question set Q and a document collection D at rank n is defined as:

$$COVERAGE(Q,D,n) \equiv \frac{\left|\left\{q \in Q \middle| R_{D,q,n} \cap A_{D,q} \neq 0\right\}\right|}{|Q|} \quad (4)$$

Coverage gives the proportion of the question set for which a correct answer can be found within the top n documents retrieved for each question.

The figure 2 shows the coverage results on Spanish. It compares our n-gram model against the vector space model. From the figure, it is possible to appreciate the improvement of our model with respect to the classical vector model. This improvement was slightly greater for passages of one sentence, but it was also noticed when using passages of three sentences.

We can also observe that the bigger the size of the passage, the greater the resultant coverage. We believe this situation is produced by some anaphoric phenomena. It indicates that the answer is not always located in the sentence containing the n-grams of the question, but in the previous or following sentences. However, even when the bigger passages produce better coverage results, the small passages are preferred. This is because the complexity of the answer extraction (next module in the QA process) increases when dealing with bigger passages.

[1] The Cross-Language Evaluation Forum; http://clef.iei.pi.cnr.it/

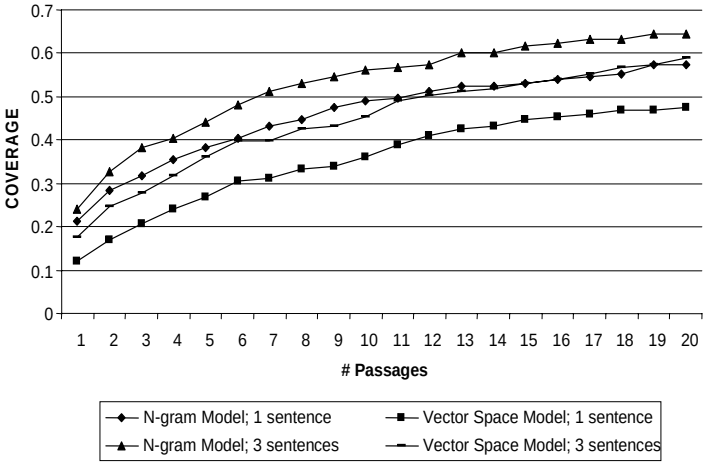

Figure 2. Comparison against the vector space model

The figure 3 shows the coverage results on Spanish, Italian and French. These results were obtained considering passages of three sentences. It is important to notice that our *n*-gram PR model is very stable on the three different languages. In all the cases, the coverage was superior to 60% for the first twenty passages. The small differences favoring the Spanish experiment could be produced because of the size, and the possible redundancy, of the collection (see table 1).

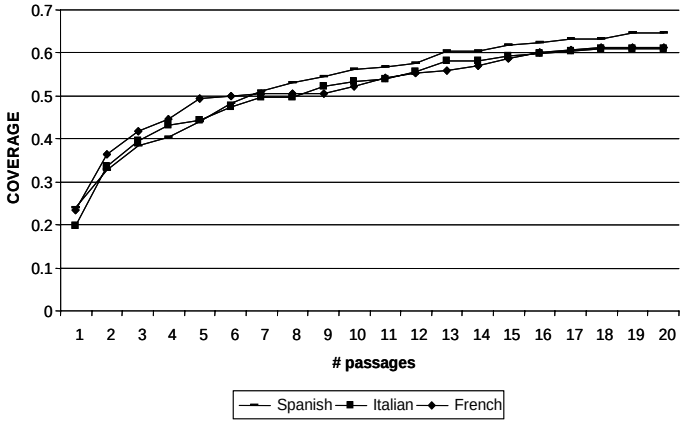

Figure 3. Coverage on Spanish, Italian and French

Another important characteristic of our model is the high redundancy of the correct answers. The figure 4 indicates that the correct answer occurs in average four times among the top twenty passages. This finding is very important since it makes our system suitable for those current answer extraction methods based on statistical approaches [4, 13, 14, 16, 3, 5, 17].

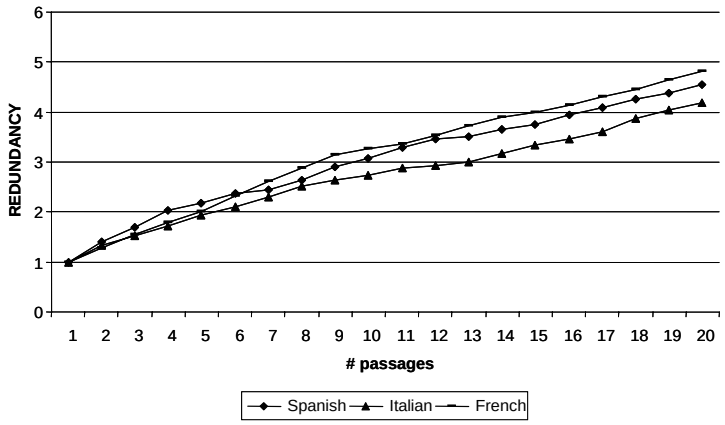

Figure 4. Redundancy on Spanish, Italian and French

4 Conclusions

Passage Retrieval (PR) is commonly used as the first step in current QA systems. In this paper, we have proposed a new PR model based on statistical *n*-gram matching. This model, which allowed us to obtain passages that contain the answer for a given question, outperforms the classic vector space model for passage retrieval, giving a higher coverage with a high redundancy (i.e., the correct answer was found more than once in the returned passages).

Moreover, this PR model does not make use of any linguistic information and thus it is almost language independent. The experimental results on Spanish, Italian and French confirm this feature and show that the proposed model is stable for different languages.

As a future work we plan to study the influence of the size and redundancy of the document collection on the coverage results. Our intuition is that the proposed model is more adequate for very large document collections.

In addition, we consider that this model should allow to tackle the problem of the Multilingual QA since it will be able to distinguish what translations are better looking for their *n*-gram structure in the corpus, and it will discriminate the bad translations as it is very unlikely that they appear. Our further interest is to proof the above assumption using as input several automatic translations and merging the returned passages. Those passages obtained with bad translations will have less weight than those that correspond to the correct ones.

Acknowledgements

We would like to thank CONACyT for partially supporting this work under the grant 43990A-1 as well as R2D2 CICYT (TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects.

References

1. Corrada-Emanuel, A., Croft, B., Murdock, V.: Answer passage retrieval for question answering. Technical Report, Center for Intelligent Information Retrieval (2003).
2. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Multilingual question/answering the DIOGENE system. In: 10th Text Retrieval Conference (2001).
3. Aunimo, L., Kuuskoski, R., Makkonen, J.: Cross-language question answering at the University of Helsinki. In: Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004), Bath, UK (2004).
4. Vicedo, J.L., Izquierdo, R., Llopis, F., Muñoz, R.: Question answering in Spanish. In: Workshop of the Cross-Lingual Evaluation Forum (CLEF 2003), Trondheim, Norway (2003).
5. Neumann, G., Sacaleanu, B.: Experiments on robust nl question interpretation and multi-layered document annotation for cross-language question/answering system. In: Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004), Bath, UK (2004).
6. Hovy, E., Gerber, L., Hermjakob, U., Junk, M., Lin, C.: Question answering in webclopedia. In: Ninth Text Retrieval Conference (2000).
7. Roberts, I., Gaizauskas, R.J.: Data-intensive question answering. In: ECIR. Lecture Notes in Computer Science, Vol. 2997, Springer (2004).
8. Gaizauskas, R., Greenwood, M.A., Hepple, M., Roberts, I., Saggion, H., Sargaison, M.: The university of Sheffield´s TREC 2003 Q&A experiments. In: The 12th Text Retrieval Conference (2003).
9. Greenwood, M.A.: Using pertainyms to improve passage retrieval for questions requesting information about a location. In: SIGIR (2004).
10. Ahn, R., Alex, B., Bos, J., Dalmas, T., Leidner, J.L., Smillie, M.B.: Cross-Lingual question answering with QED. In: Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004), Bath, UK (2004).
11. Hess, M.: The 1996 international conference on tools with artificial intelligence (tai´96). In SIGIR (1996).
12. Liu, X., Croft, W.: Passage retrieval based on language models (2002).
13. Del-Castillo-Escobedo, A., Montes-y-Gómez, M., Villaseñor-Pineda, L.: QA on the Web: a preliminary study for spanish language. In: Proceedings of the fifth Mexican International Conference in Computer Science (ENC´04), Colima, Mexico (2004).
14. Brill, E., Lin, J., Banko, M., Dumais, S.T., Ng, A.Y.: Data-intensive question answering. In: 10th Text Retrieval Conference (2001).
15. Buchholz, S.: Using grammatical relations, answer frequencies and the world wild web for trec question answering. In: 10th Text Retrieval Conference (2001).
16. Brill, E., Dumais, S., Banko, M.: An analysis of the askmsr question answering system (2002).
17. Costa, L.: First evaluation of esfinge: a question answering system for Portuguese. In: Workshop of the Cross-Lingual Evaluation Forum (CLEF 2004), Bath, UK (2004).