A Full Data-Driven System for Multiple Language Question Answering

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

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Mexico.
{mmontesg, villasen, mapco}@inaoep.mx
[2] Universidad Politécnica de Valencia (UPV), Spain.
{jogomez,esanchis, prosso}@dsic.upv.es

Abstract. This paper describes a full data-driven system for question answering. The system uses pattern matching and statistical techniques to identify the relevant passages as well as the candidate answers for factoid and definition questions. Since it does not consider any sophisticated linguistic analysis of questions and answers, it can be applied to different languages without requiring major adaptation changes. Experimental results on Spanish, Italian and French demonstrate that the proposed approach can be a convenient strategy for monolingual and multilingual question answering.

1 Introduction

The amount of documents available online is increasing every day. As a consequence, better information retrieval methods are required to achieve the needed information. Question Answering (QA) systems are information retrieval applications whose aim is to provide inexperienced users with a flexible access to the information. These systems allow users to write a query in natural language and to obtain not a set of documents which contain the answer, but the concise answer itself [9]. That is, given a question like: "Where is the Popocatepetl located?", a QA system must respond "Mexico", instead of just returning a list of documents related to the volcano.

Recent developments in QA use a variety of linguistic resources to help in understanding the questions and the documents. The most common linguistic resources include: part-of-speech taggers, parsers, named entity extractors, dictionaries, and WordNet [1,5,6]. Despite promising results, these approaches have two main inconveniences: (i) the construction of such linguistic resources is very complex; and (ii) these resources are highly binding to a specific language.

In this paper we present a QA system that allows answering factoid and definition questions. This system is based on a full *data-driven approach* [2], which requires minimum knowledge about the lexicon and the syntax of the specified language. Mainly, it is supported by the idea that the questions and their answers are commonly

---

[*] This work is a revised version of the paper "INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering", previously published in the CLEF 2005 working notes (www.clef-campaign.org/2005/working_notes/).

expressed using the same set of words, and therefore, it simply uses a lexical pattern matching method to identify relevant document passages and to extract the candidate answers.

The proposed approach has the advantage of being adaptable to several different languages, in particular to moderately inflected languages such as Spanish, English, Italian and French. Unfortunately, this flexibility has its price. To obtain a good performance, the approach requires the use of a redundant target collection, that is, a collection in which the answers to questions occur more than once. On one hand, this redundancy increases the probability of finding a passage containing a simple lexical matching between the question and the answers. On the other hand, it enhances the answer extraction, since correct answers tend to be more frequent than incorrect responses.

The proposed system also uses a set of heuristics that attempt to capture some regularities of language and some stylistic conventions of newsletters. For instance, it considers that most named entities are written with an initial uppercase letter, and that most concept definitions are usually expressed using a very small number of fixed arrangements of noun phrases. This kind of heuristics guides the extraction of the candidate answers from the relevant passages.

2 System Overview

Figure 1 shows the general architecture of our system, which is divided into two main modules. One focuses on answering factoid questions. It considers the tasks of: (i) *passage indexing*, where documents are preprocessed, and a structured representation of the collection is built; (ii) *passage retrieval*, where the passages with the greatest probability to contain the answer are recovered from the index; and (iii) *answer extraction*; where candidate answers are ranked and the final answer recommendation of the system is produced.

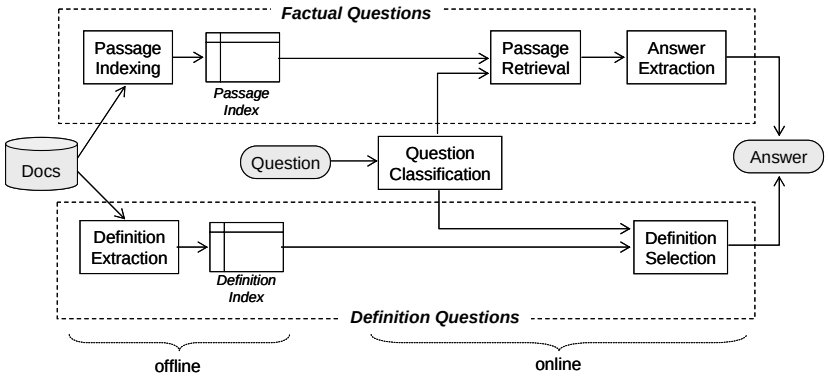

Figure 1. Block diagram of the system

The other module concentrates on answering definition questions. It includes the tasks of: (i) *definition extraction*; where all possible pairs of acronym-meaning and person-position[1] are located and indexed; and (ii) *definition selection*, where the relevant data pairs are identified and the final answer of the system is generated.

The following sections describe in detail these modules.

3 Answering Factoid Questions

3.1 Passage Retrieval

The Passage Retrieval (PR) method is specially suited for the QA task [4]. It allows retrieving the passages with the highest probability to contain the answer, instead of simply recovering the passages sharing a subset of words with the question.

Given a user question, the PR method finds the passages with the relevant terms (non-stopwords) using a classical information retrieval technique based on the vector space model. Then, it 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 largest *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, it returns to the user the passages with the new weights.

3.1.1 Similarity Measure

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

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

Where $sim(d, q)$ is a function which measures the similarity of the set of *n*-grams of the passage *d* with the set of *n*-grams of the question *q*. D_j is the set of *j*-grams of the passage *d* and Q_j is the set of *j*-grams that are generated from the question *q*. That is, D_1 will contain the passage unigrams whereas Q_1 will contain the question unigrams, D_2 and Q_2 will contain the passage and question bigrams respectively, and so on until D_n and Q_n. In both cases, *n* is the number of question terms.

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

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

---

[1] In general, we consider the extraction of person-description pairs.

[2] We introduce the notation $x(n)$ for the sake of simplicity. In this case $x(n)$ indicates the *n*-gram *x* of size *n*.

$$h\left(x(j), D_j\right) = \begin{cases} \sum_{k=1}^{j} w_{\hat{x}_k(1)} & \text{if } x(j) \in D_j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where the notation $\hat{x}_k(1)$ indicates the *k*-th unigram included in the *j*-gram *x*, and $w_{\hat{x}_k(1)}$ specifies the associated weight to this unigram. This weight gives an incentive to the terms –unigrams– that appear rarely in the document collection. Moreover, this weight should also discriminate the relevant terms against those (e.g. stopwords) which occur often in the document collection.

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

$$w_{\hat{x}_k(1)} = 1 - \frac{\log\left(n_{\hat{x}_k(1)}\right)}{1 + \log(N)} \quad (3)$$

Where $n_{\hat{x}_k(1)}$ is the number of passages in which appears the unigram $\hat{x}_k(1)$, and *N* is the total number of passages in the collection. We assume that the stopwords occur in every passage (i.e., *n* 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.

3.2 Answer Extraction

This component aims to establish the best answer for a given question. In order to do that, it first determines a small set of candidate answers, and then, it selects the final unique answer taking into consideration the position of the candidate answers inside the retrieved passages.

The algorithm applied to extract the most probable answer from the given set of relevant passages is described below[3]:

1. Extract all the unigrams that satisfy some given typographic criteria. These criteria depend on the type of expected answer. For instance, if the expected answer is a named entity, then select the unigrams starting with an uppercase letter, but if the expected answer is a quantity, then select the unigrams expressing numbers.
2. Determine all the *n*-grams assembled from the selected unigrams. These *n*-grams can only contain the selected unigrams and some stopwords.
3. Rank the *n*-grams based on their compensated frequency. The compensated frequency of the *n*-gram *x*(*n*) is computed as follows:

$$F_{x(n)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n-i+1} \frac{f_{\hat{x}_j(i)}}{\sum_{\forall y \in G_i} f_{y(i)}} \quad (4)$$

where G_i indicates the set of *i*-grams, *y*(*i*) represents the *i*-gram *y*, $\hat{x}_j(i)$ is the *j*-th *i*-gram included in *x*(*n*), $f_{y(i)}$ specifies the frequency of occurrence of the *i*-gram *y*, and $f_{x(n)}$ indicates the compensated frequency of *x*(*n*).

[3] For more details please refer to (Del-Castillo et al., 2004).

4. Select the top five *n*-grams as candidate answers.
5. Compute a ranking score for each candidate answer. This score is defined as the weight of the first retrieved passage (refer to formula 1) that contains the candidate answer.
6. Select as the final response the candidate answer with the greatest ranking score. If two or more of the candidate answers have the same ranking score, then select the one with the greatest compensated frequency.

4 Answering Definition Questions

Our system uses an alternative method to answer definition questions. This method makes use of some regularities of language and some stylistic conventions of news letters to capture the possible answer for a given definition question. A similar approach was presented in [7,8].

The process of answering a definition question considers two main tasks. First, the *definition extraction*, which detects text segments containing a description of a term (in particular we consider descriptions related to person's positions and organization's acronyms). Then, the *definition selection*, where the most relevant description for a given question term is identified and the final answer of the system is generated.

4.1 Definition Extraction

The regularities of language and the stylistic conventions of news letters are captured by two basic lexical patterns. These patterns allow the construction of two different definition catalogs. The first one includes a list of pairs of acronym-meaning. The other one consists of a list of person-position couples.

In order to extract the acronym-meaning pairs we use an extraction pattern based on the use of parentheses:

$$w_1 \text{ <meaning> (<acronym>)} \tag{5}$$

In this pattern, w_1 is a lowercase non stopword, <*meaning*> is a sequence of words starting with an uppercase letter (that may also include some stopwords), and <*acronym*> indicates an uppercase single word.

By means of this pattern we could identify pairs like [*AAPNP – la Asociación de Armadores de Pesca del Norte Portugués*]. In particular, this pair was extracted from the following paragraph:

> *"El pasado 3 de enero la Asociación de Armadores de Pesca del Norte Portugués (AAPNP) acusó al ministro de Asuntos Marítimos, Eduardo Azevedo Soares, de favorecer a los pesqueros españoles."*

In contrast, the extraction of person-position pairs is guided by the occurrence of a special kind of appositive phrase. This information is encapsulated in the following extraction pattern.

$$w_1\ w_2 \text{ <description> , <referent> [,|.]} \tag{6}$$

Where w_1 represents any word, except for the prepositions "of" and "in", w_2 is an article, <*description*> is a free sequence of words, and <*referent*> indicates a sequence of words starting with an uppercase letter.

Applying this extraction pattern over the below paragraph we caught the pair [*Manuel Concha Ruiz – jefe de la Unidad de Trasplantes del hospital cordobés*].

> "*… no llegó a Córdoba hasta las 19:30 horas, se llevó a cabo con un corazón procedente Madrid y fue dirigida por el jefe de la Unidad de Trasplantes del hospital cordobés, Manuel Concha Ruiz.*"

4.2 Definition Selection

The main quality of the above extraction patterns is their generality: they can be applied to different languages without requiring major adaptation changes. However, this generality causes the patterns to often extract non-relevant information, i.e., information that does not indicate an acronym-meaning or person-position relation. For instance, when applying the pattern (6) to the next text segment we identified the incorrect pair [*Manuel H. M. – otros dos pasajeros de este vehículo*].

> *"También el conductor del Opel Corsa, Antonio D.V., de 24 años, y los ocupantes Andrés L.H., de 24, y Francisco F.L, de 21, resultaron con heridas graves, mientras que los otros dos pasajeros de este vehículo, Manuel H.M., de 29, y Miguel J.M., de 25, resultaron con heridas leves".*

Since the catalogs contain a mixture of correct and incorrect definition pairs, it is necessary to do an additional process in order to select the most probable answer for a given definition question. This process is supported on the idea that the correct information is more redundant than the incorrect one. It considers the following two criteria:

1. The most frequent definition in the catalog has the highest probability to be the correct answer.
2. The larger and, therefore, more specific definitions tend to be the more pertinent answers.

In order to increase the opportunity of selecting the correct answers, the definition catalogs must be cleaned before the execution of this process. We consider two main actions: (i) the removal of stopwords at the beginning of descriptions –acronym meanings and person positions; and (ii) the elimination of the acronym meanings having fewer words than letters in the acronym.

The following example illustrates the selection process. Assume that the user question is "*who is Manuel Conde?*", and that the definition catalog contains the records shown below. Then, the method selects the description "*presidente de la Comisión de Paz del Parlamento Centroamericano (PARLACEN)*" as the most probable answer.

> *Manuel Conde: gobierno de Serrano*
> *Manuel Conde: gobierno de Jorge Serrano (1991-1993)*
> *Manuel Conde: gobierno de Jorge Serrano*
> *Manuel Conde: ex presidente de la COPAZ que participó en la primera etapa*

Manuel Conde: presidente de la Comisión de Paz del Parlamento Centroamericano (PARLACEN)
Manuel Conde: presidente de la Comisión de Paz del Parlamento Centroamericano (PARLACEN)

5 Evaluation Results

This section presents the evaluation results of our system at the QA@CLEF2005 monolingual tracks for Spanish, Italian and French. In the three languages, the evaluation exercise consisted of answering 200 questions of three basic types: factoid, definition and temporal restricted. In all cases, the target corpora were collections of news articles. Table 1 shows some general numbers on the evaluation data set.

Table 1. The evaluation data set

	Target corpora	Question set		
	# sentences	Factoid	Definition	Temporal
Spanish	5,636,945	118	50	32
Italian	2,282,904	120	50	30
French	2,069,012	120	50	30

Figure 2 shows our global results on the three languages[4]. The Spanish results were better than those for Italian and French. However, we obtained the best evaluation result in Italian. In this case the average precision was of 24.1%. In the monolingual Spanish and French tasks we achieved the second best results. In Spanish, the best result was of 42% and the average precision of 31.7%. In French, the best preci-

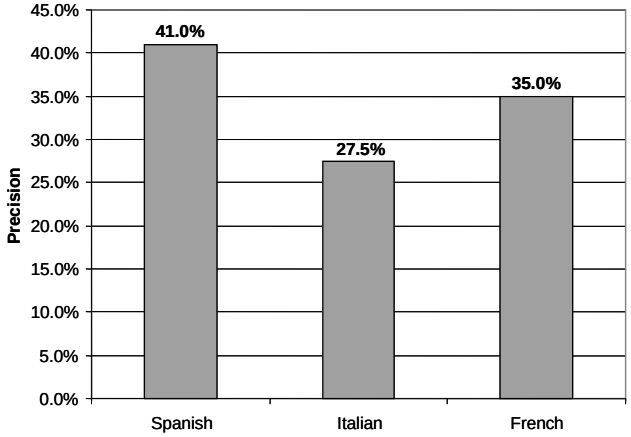

Figure 2. Overall accuracy results

[4] Since our system only distinguishes between factoid and definition questions, we treated the temporal-restricted questions as simple factoid.

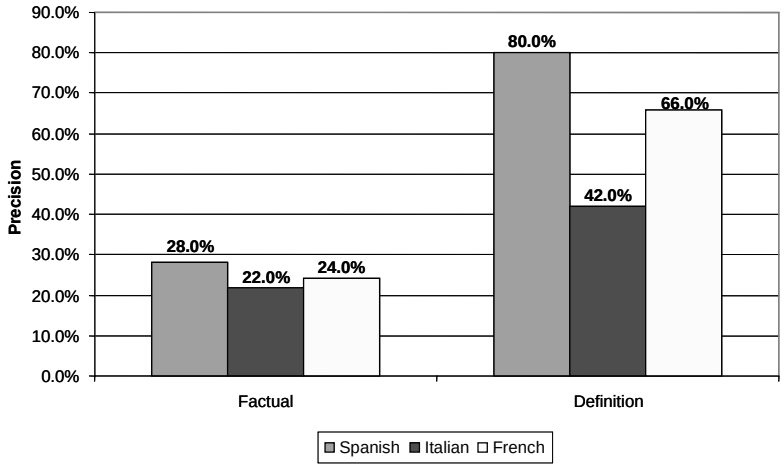

Figure 3. Accuracy on factoid and definition questions

sion was of 64%, and the average of 34%.

Figure 3 detail our results by question types. It can be noticed that we are significantly better in answering definition questions. However, the numbers indicate that the method for answering factoid questions is language independent, while the approach for answering definition questions tends to be more language dependent.

Conclusions

This paper presented a question-answering system based on a full *data-driven approach*. The system is supported by the idea that the questions and their answers are commonly expressed using the same words, and therefore, it simply uses pattern matching and statistical techniques to identify the relevant passages as well as the candidate answers for factoid and definition questions.

The experiments on Spanish, Italian and French showed the potential and portability of our approach. They also indicated that our method for answering factoid question, which is based on the matching and counting of *n*-grams, is *language-independent*. However, it greatly depends on the redundancy of the answers in the target collection. On the contrary, the method for answering definition questions is very precise. Nevertheless, we cannot conclude anything about its language independence.

Futere work includes improving the ranking score for factoid questions, in order to reduce the dependence on the data redundancy. We also plan to design a technique to discover extraction patterns on the Web. This will help in decreasing the language dependence of our method for answering definition questions.

Acknowledgements. This work was done under partial support of CONACYT (Project Grant 43990), R2D2 (CICYTTIC2003-07158-C04-03), and ICT EU-India (ALA/95/23/2003/077-054). We would also like to thank the CLEF organization committee.

References

1. Ageno, A., Ferrés, D., González, E., Kanaan, S., Rodríguez H., Surdeanu, M., and Turmo, J. *TALP-QA System for Spanish at CLEF-2004*. Working Notes for the CLEF 2004 Workshop, Bath, UK, 2004.
2. Brill, E., Lin, J., Banko, M., Dumais, S., and Ng, A. *Data-intensive Question Answering.* TREC 2001 Proceedings, 2001.
3. Del-Castillo, A., Montes-y-Gómez, M., and Villaseñor-Pineda, L. *QA on the web: A preliminary study for Spanish language*. Proceedings of the 5th Mexican International Conference on Computer Science (ENC04), Colima, Mexico, 2004.
4. Gómez-Soriano, J.M., Montes-y-Gómez, M., Sanchis-Arnal, E., Rosso P. *A Passage Retrieval System for Multilingual Question Answering*. 8th International Conference on Text, Speech and Dialog, TSD 2005. Lecture Notes in Artificial Intelligence, vol. 3658, 2005.
5. Jijkoun, V., Mishne, G., de Rijke, M., Schlobach, S., Ahn, D., and Müller, K.. *The University of Amsterdam at QA@CLEF 2004*. Working Notes for the CLEF 2004 Workshop, Bath, UK, 2004.
6. Pérez-Coutiño, M., Solorio, T., Montes-y-Gómez, M., López-López, A., and Villaseñor-Pineda, L. *The Use of Lexical Context in Question Answering for Spanish*. Working Notes for the CLEF 2004 Workshop, Bath, UK, 2004.
7. Ravichandran D. and Hovy E. *Learning Surface Text Patterns for a Question Answering System*. In ACL Conference, 2002.
8. Saggion, H. *Identifying Definitions in Text Collections for Question Answering*. LREC 2004.
9. Vicedo, J.L., Rodríguez, H., Peñas, A. and Massot, M. *Los sistemas de Búsqueda de Respuestas desde una perspectiva actual*. Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, n.31, 2003.