INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering

M. Montes-y-Gómez[1], L. Villaseñor-Pineda[1], M. Pérez-Coutiño[1]
J. M. Gómez-Soriano[2], E. Sanchis-Arnal[2], and P. Rosso[2]

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

Abstract. Recent works on question answering are based on complex natural language processing techniques: named entity extractors, parsers, chunkers, etc. While these approaches have proven to be effective they have the disadvantage of being targeted to a particular language. In this paper we present a full data-driven method that uses simple lexical pattern matching and statistical techniques in order to identify the relevant passages as well as the more probable candidate answers for factual and definition questions. The main quality of this method is that it can be applied to different languages without requiring major adaptation changes. Experimental results of the method in Spanish, Italian and French show that the approach can be a practical solution for monolingual and multilingual question answering applications.

1 Introduction

The volume of online available documents is growing 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, allowing them writing a query in natural language and obtaining not a set of documents which contain the answer, but the concise answer itself (Vicedo et al, 2003). 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 (Jijkoun et al., 2004; Ageno et al., 2004; Pérez-Coutiño et al., 2004). Despite of the promising results of these approaches, they have two main inconveniences: (i) the construction of such linguistic resources is a very complex task; and (ii) these resources are highly binding to a specific language.

In this paper we present a QA system that allows answering factual and definition questions. This system is based on a full *data-driven approach* (Brill et al., 2001), which requires minimum knowledge about the lexicon and the syntax of the specified language. Mainly, it is supported on the idea that the questions and their answers are commonly 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 to be easily adapted to several different languages, in particular to moderately inflected languages such as Spanish, English, Italian and French. Unfortunately, this generality has its price. To obtain a good performance, the approach requires using a redundant target collection, that is, a collection in which the question answers occurs more than once. On the 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 presented system also uses a set of heuristics that attempt to capture some regularities of language and some stylistic conventions of news letters. 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.

In the rest of the paper we present the main architecture of our data-driven QA system. We also discuss the evaluation results on Spanish, Italian and French.

2 System Overview

The figure 1 shows the general architecture of our system. It is divided in two main modules. One of them focuses on answering factual questions. It considers the tasks of *passage indexing*, where documents are preprocessed, and a structured representation of the collection is built; *passage retrieval*, where the passages with more probability to contain the answer are recovered from the index; and *answer extraction*; where candidate answers are ranked and the final answer recommendation of the system is produced.

The other module concentrates on answering definition questions. It includes the tasks of *definition extraction*; where all possible pairs of acronym-meaning and referent-description are located and indexed; and *definition selection*, where the relevant data pair is identified and the final answer of the system is generated.

The following sections describe each of these modules and their main tasks.

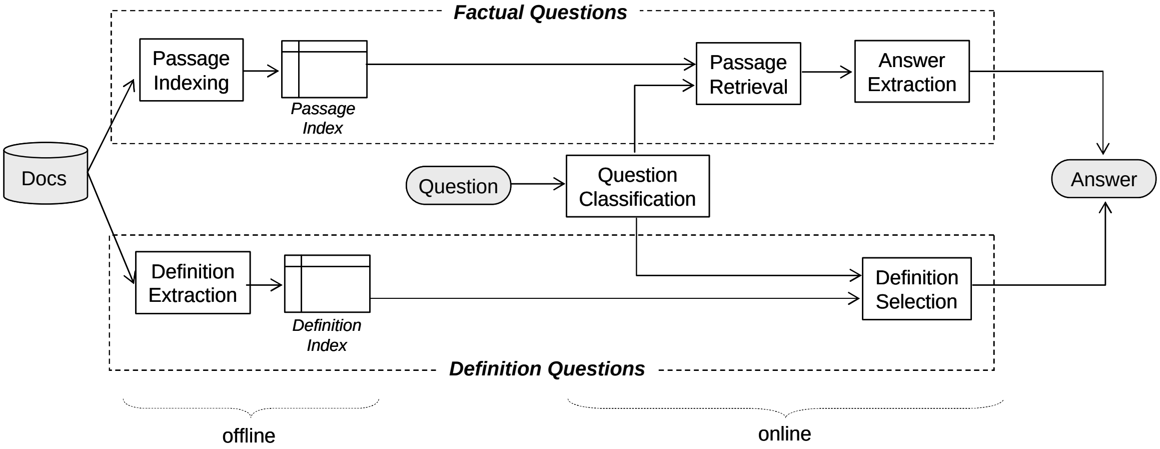

Figure 1. Block diagram of the system

3 Answering Factual Questions

3.1 Passage Retrieval

The passage retrieval (PR) method is specially suited for the QA task (Gómez-Soriano et al., 2005). It allows retrieving the passages with the highest probability to contain the answer, instead of simply recover 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).

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

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*. 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. 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 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[1]. The function h assigns a weight to every question n-gram as defined in (2).

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

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 often occur in the document collection.

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

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

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 better answer to the 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 more probable answer from the given set of relevant passages is described below. For more detail refer to (Del-Castillo et al., 2004).

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 we select the unigrams starting with an uppercase letter. But if the answer must be a quantity, then we 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:

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

   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 $x(n)$.
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 respond the candidate answer with the greatest ranking score. In the case that two or more of the candidate answers have the same ranking score, then we select the one with the greatest compensated frequency.

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

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 (Ravichandran et al., 2001; Saggion, 2004).

The process of answering a definition question considers to main tasks. First, the *definition extraction*, which detects the text segments that contains the description or meaning of a term (in particular those related with the name of a person or an organization). 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 language regularities and the stylistic conventions of news letters are captured by two basic lexical patterns. These patterns allow constructing two different definition catalogs. The first one includes a list of pairs of acronym-meaning. The other one consists of a list of referent-description couples.

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

$$w_1 \textit{<meaning>} \ (\ \textit{<acronym>} \ ) \qquad (5)$$

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

By means of this pattern we could identify pairs like [*PARM – Partido Auténtico de la Revolución Mexicana*]. In particular this pair was catch from the following paragraph:

> "*El Partido Auténtico de la Revolución Mexicana (PARM) nombró hoy, sábado, a Álvaro Pérez Treviño candidato presidencial de ese organismo para las elecciones federales del 21 de agosto de 1994*".

In contrast, the extraction of referent-description pairs is guided by the occurrence of a special kind of appositive phrases. This information was encapsulated in the following extraction pattern.

$$w_1 \ w_2 \ \textit{<description>} \ , \ \textit{<referent>} \ , \qquad (6)$$

Where w_1 may represent any word, except for a preposition, w_2 is a determiner, <*description*> is a free sequence of words, and <*referent*> indicates a sequence of words starting with an uppercase letter or being in the stopwords list.

Applying this extraction pattern over the below paragraph we could find the pair [*Alain Lombard - El director de la Orquesta Nacional de Burdeos*].

> "*El director de la Orquesta Nacional de Burdeos, Alain Lombard, ha sido despedido por el Ayuntamiento de esta ciudad, que preside Alain Juppé, tras un informe que denuncia malos funcionamientos y gastos excesivos*".

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 a relation acronym-meaning or referent-description. For instance, when using the extraction pattern (5) to analyze the following news we obtain the incorrect definition pair [Ernie Els - AFS]. In this case the resultant pair does not express an acronym-meaning relation; instead it indicates a person-nationality association.

> *Ernie Els (AFS) se mantiene en cabeza de la lista de ganancias de la "Orden de Mérito" de golf, con más de 17 millones de pesetas, mientras que el primer español es Miguel Angel Martín, situado en el puesto decimoséptimo, con 4.696.020.*

Given that the catalogs contains a mixture of correct and incorrect relation pairs, it is necessary to do an additional process in order to select the most probable answer for a given definition question. The proposed approach is supported on the idea that, on the one hand, the correct information is more abundant than the incorrect one, and on the other hand, that the correct information is redundant.

Thus, the process of definition selection considers the following two criteria:

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

The following example illustrates the process. Assume that the user question is "*who is Félix Ormazabal?*", and that the definition catalog contains the records showed below. Then, the method selects the description "*diputado general de Alava*" as the most probable answer. This decision is based on the fact that this answer is the more frequent description in the catalog related to Félix Ormazabal.

Félix Ormazabal: Joseba Egibar:
Félix Ormazabal: candidato alavés:
Félix Ormazabal: diputación de este territorio:
Félix Ormazabal: presidente del PNV de Alava y candidato a diputado general:
Félix Ormazabal: nuevo diputado general
Félix Ormazabal: diputado Foral de Alava
Félix Ormazabal: través de su presidente en Alava
Félix Ormazaba : diputado general de Alava
Félix Ormazabal: diputado general de Alava
Félix Ormazabal: diputado general de Alava

5 Evaluation Results

We participate in the evaluation task on three different languages: Spanish, Italian and French. For each language we submitted two runs. The first group of them (*tova051* runs) implements the system as described in the previous sections. The second group (*tova052* runs) resolves some factual questions as if they were definition questions. The selected questions were those asking for the name of a personality or for the acronym of an organization.

Table 1 shows our global results on the three languages. It is noticed that the Spanish results were slightly better than the Italian and French ones.

Table 1. Overall accuracy results

	tova051itit	tova052itit	tova051frfr	tova052frfr	tova051eses	tova052eses
Right	53	55	69	70	82	77
Wrong	138	135	121	120	109	113
Inexact	9	10	10	10	7	8
Unsupported	0	0	0	0	2	2
Overall Accuracy	26.5%	27.5%	34.5%	35.0%	41.0%	38.5%

The following tables detail our results by question types. Table 2 shows the accuracy on factual questions; table 3 indicates the results on definition questions, and finally, table 4 shows the achieved results on temporal questions. Our general conclusion is that the method for answering factual questions is language independent. Unfortunately we can assert the same for our approach to answer definition questions.

On the other hand, it is important to mention that the temporal questions were treated as if they were factual questions. Currently we do not have a specific method for answering this kind of questions.

Table 2. Accuracy on factual questions

	tova051itit	tova052itit	tova051frfr	tova052frfr	tova051eses	tova052eses
Right	26	28	32	33	34	28
Wrong	89	86	84	83	77	82
Inexact	5	6	4	4	5	6
Unsupported	0	0	0	0	2	2
Accuracy	21.7%	23.3%	26.7%	27.5%	28.8%	23.7%

Table 3. Accuracy on definition questions

	tova051itit	tova052itit	tova051frfr	tova052frfr	tova051eses	tova052eses
Right	21	21	33	33	40	40
Wrong	26	26	12	12	9	9
Inexact	3	3	5	5	1	1
Unsupported	0	0	0	0	0	0
Accuracy	42.0%	42.0%	66.0%	66.0%	80.0%	80.0%

Table 4. Accuracy on temporal questions

	tova051itit	tova052itit	tova051frfr	tova052frfr	tova051eses	tova052eses
Right	6	6	4	4	8	9
Wrong	23	23	25	25	23	22
Inexact	1	1	1	1	1	1
Unsupported	0	0	0	0	0	0
Accuracy	18.8%	18.8%	12.5%	12.5%	26.7%	30.0%

As we mention in section 4, the definition catalogs have several erroneous registers. Also, they are incomplete, since they do not include all possible acronym-meaning and referent-description pairs. Nevertheless, they contain a large amount of registers and constitute a valuable information repository for answering definition questions. The tables 5 and 6 compare the catalogs that were extracted for each language.

Table 5. Data of the acronym-meaning catalog

	Acronym-meaning catalog			Document collection			
	#acronyms	#meanings	meanings per acronym	# sentences	#words	acronyms per sentence	meanings per sentences
Spanish	14,921	263,388	17.65	5,636,945	151,553,838	0.0026	0.046
French	8,375	66,690	7.96	2,069,012	45,057,929	0.0040	0.032
Italian	4,588	21,606	4.71	2,282,904	49,343,596	0.0020	0.009

Table 6. Data of the referent-description catalog

	Referent-description catalog			Document collection			
	#referents	#descriptions	descriptions per referent	#sentences	#words	referents per sentence	desriptions per sentence
Spanish	131,356	563,411	4.29	5,636,945	151,553,838	0.02330	0.09995
French	31,864	58,905	1.85	2,069,012	45,057,929	0.01540	0.02847
Italian	45,856	125,023	2.73	2,282,904	49,343,596	0.02009	0.05476

The above tables reveal an important association between the sizes of the document collection and the generated catalogs. This is of great relevance since our approach of answer selection is mainly based on the definition redundancy. However, the tables also seem indicate that the extraction patterns are not totally language independent. This assertion is based on the fact that we obtain fewer descriptions per referent for the French, even when the Italian collection is slightly smaller.

The tables 7 and 8 show our results on answering definition questions related to acronyms as well as to personality descriptions.

Table 7. Accuracy on questions about acronyms

	tova051eses	tova052eses	tova051frfr	tova052frfr	tova051itit	tova052itit
Right	20	20	14	14	10	10
Wrong	8	8	11	11	22	22
ineXact	1	1	3	3	2	2
Unsopported	0	0	0	0	0	0
Accuracy	69.0%	69.0%	50.0%	50.0%	29.4%	29.4%

Table 8. Accuracy on questions about personality descriptions

	tova051eses	tova052eses	tova051frfr	tova052frfr	tova051itit	tova052itit
Right	20	20	19	19	11	11
Wrong	1	1	1	1	4	4
ineXact	0	0	2	2	1	1
Unsopported	0	0	0	0	0	0
Acurracy	95.2%	95.2%	86.4%	86.4%	68.8%	68.8%

6 Conclusions

This paper presents a question answering system that allows answering factual and definition questions. This system is based on a full *data-driven approach*. The main idea behind the approach is that the questions and their answers are commonly 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 experiments on Spanish, Italian and French have shown the potential and portability of our approach. They also indicated that our method for answering factual question, which is based on the matching and counting of *n*-grams, is *language independent*. However, this method greatly depends on the redundancy of the answers in the target collection. This condition limited the method to a poor accuracy.

On the contrary, the method for answering definition questions is very precise. Nevertheless, we can not conclude about it language independence.

As future work we plan to improve the ranking score for the factual answers. This will help in reducing the dependence of our method to the data redundancy. We also consider to evaluate the quality of the definition catalogs in order to conclude something about the language independence of our approach.

Acknowledgements. This work was done under partial support of CONACYT (Project Grant 43990), SNI-Mexico, and the Human Language Technologies Laboratory of INAOE. We also like to thanks to 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*. To appear in the proceedings of the 8th International Conference on Text, Speech and Dialog, TSD 2005. Karlovy Vary, Czech Republic, 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.