Toward a Document Model for Question Answering Systems

M. Pérez-Coutiño, T. Solorio, M. Montes-y-Gómez[†], A. López-López and L. Villaseñor-Pineda

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luis Enrique Erro No. 1, Sta Ma Tonantzintla, 72840, Puebla, Pue, México.
{mapco,thamy,mmontesg,allopez,villasen}@inaoep.mx

Abstract. The problem of acquiring valuable information from the large amounts available today in electronic media requires automated mechanisms more natural and efficient than those already existing. The trend in the evolution of information retrieval systems goes toward systems capable of answering specific questions formulated by the user in her/his language. The expected answers from such systems are short and accurate sentences, instead of large document lists. On the other hand, the state of the art of these systems is focused –mainly– in the resolution of factual questions, whose answers are named entities (dates, quantities, proper nouns, etc). This paper proposes a model to represent source documents that are then used by question answering systems. The model is based on a representation of a document as a set of named entities (NEs) and their local lexical context. These NEs are extracted and classified automatically by an off-line process. The entities are then taken as instance concepts in an upper ontology and stored as a set of DAML+OIL resources which could be used later by question answering engines. The paper presents a case of study with a news collection in Spanish and some preliminary results.

Keywords: Question Answering, Ontology, Semantic Web, Named Entity Classification.

1 Introduction

The technological advances have brought us the possibility to access large amounts of information automatically, either in the Internet or in specialized collections of information. However, such information becomes useless without the appropriate mechanisms that help users to find the required information when they need it. Traditionally, searching information in non-structured or semi-structured sources been performed by search engines that return a ranked list of documents containing all or some of the terms from the user's query. Such engines are incapable of returning a concise answer to a specific information request [4].

[†] This work was done while visiting the Dept. of Information Systems and Computation Polytechnic University of Valencia, Spain.

The alternative to information retrieval systems for resolving specific questions are Question Answering (QA) systems capable of answer questions formulated by the user in natural language. Research in QA has increased as a result of the inclusion of QA evaluations as part of the Text Retrieval Conference (TREC)[1] in 1999, and recently [5] in Multilingual Question Answering as part of the Cross Language Evaluation Forum (CLEF)[2].

The goal of QA systems is to respond to a natural language question stated by the user, replying with a concrete answer to the given question and, in some cases, a context for its validation. Current operational QA systems are focused in factual questions [1, 15] that require a named entity (date, quantity, proper noun, locality, etc) as response. For instance, the question *"¿Dónde nació Benito Juárez?"*[3] demands as answer *'San Pablo Guelatao'*, a locality of Mexico. Several approaches of QA systems like [8, 14] use named entities at different degree of refinement in order to find a candidate answer. Other systems like [3, 9] include the use of ontologies and contextual patterns of named entities to represent knowledge about question and answer contents. Thus, it is clear that named entities identification plays a central role in the resolution of factual questions.

On this basis, we propose in this paper a model for the representation of the source documents that are then used by QA systems. The proposed model represents text documents as a set of classified named entities and their local lexical context (nouns and verbs). The representation is automatically gathered by an off-line process that generates instances of concepts from a top level ontology and stores them as resources in DAML+OIL.

The rest of this paper is organized as follows; section two describes the proposed model, both at conceptual and implementation level; section three details the process of the named entities extraction and classification; section four presents a case of study answering some questions employing the model in a set of news documents in Spanish; finally section five exposes our conclusions and discusses further work.

2 Model Description

The aim of modeling source documents for QA systems is to provide a preprocessed set of resources which contain valuable information that makes easier to accomplish answer retrieval and extraction tasks. An important feature of the proposed model is that implies a uniform format for data sources, as mentioned in [1] "…it is also necessary that the data sources become more heterogeneous and of larger size…" Developing a document model makes possible that several heterogeneous data sources can be expressed in a standardized format, or at least feasible the transformation and mapping between equivalent sources.

To reach these goals, the following key assumptions were made to develop the proposed model:

1. The collection of documents that will be used by the QA system contains documents about facts like those published in news without domain restriction.

[1] http://trec.nist.gov/

[2] http://clef-qa.itc.it/

[3] Where was Benito Juarez born?

2. The model must reuse an upper ontology in order to allow further refinement and reasoning on the named entities.
3. The model must be encoded in some ontological language for the Semantic Web in order to allow future applications such as specialized QA engines or web agents that make use of the document representation, instead of the document itself, to achieve their goals.

The next subsection details the conceptual and implementation levels of the model.

2.1 Conceptual Level

Figure 1 shows the model. At the conceptual level, a document is seen as a factual text object whose content refers to several named entities even when it is focused on a central topic. Named entities could be one of these objects: persons, organizations, locations, dates and quantities. The model assumes that the named entities are strongly related to their lexical context, especially to nouns (subjects) and verbs (actions). Thus a document can be seen as a set of entities and their contexts. Moreover, each named entity could be refined by means of ontologies [6]. This is the aim of instantiating an upper level ontology, instead of developing an ontology from scratch.

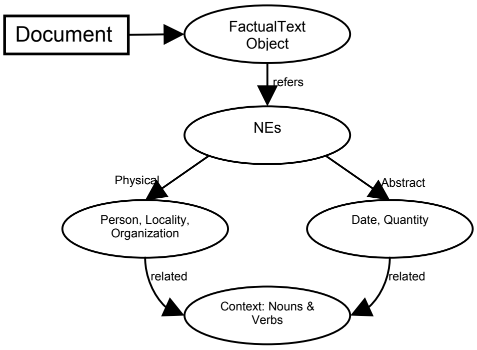

Figure 1. The proposed model.

The model is based on the Suggested Upper Merged Ontology (SUMO)[4] [7], an existing framework specifically designed to provide a basis for more specific domain ontologies. SUMO combines a number of top-level ontologies to achieve wide conceptual coverage, has a strong basis of semiotic and linguistic concepts already, and is being developed by an IEEE working group[5] that includes a number of experts from a variety of fields.

2.2 Implementation Level

As mentioned earlier the model is implemented as a set of instances of concepts in SUMO. The mapping between NEs and SUMO, as well as the used slots or axioms are shown in table 1.

[4] http://ontology.teknowledge.com:8080/

[5] http://suo.ieee.org/

The use of “refers” and “cooccurs” slots allow to refine the mapping of concepts between NEs and SUMO concepts. For instance, with an improved version of the extraction process, we could refer to a “City” instead of a “GeographicArea”, or to a “Government” instead of an “Organization”.

Table 1. Mapping between NEs and SUMO concepts.

NEs	SUMO Concept	Slot	Description
	FactualText	refers	This is the top concept of the model. Refers slot means that a factual text could make reference to other entities (like our NEs).
Person	Human	cooccurs	Human, Organization and GeographicArea could be in co-occurrence with other entities (like verbs and nouns).
Organization	Organization	cooccurs	
Locality	GeographicArea	cooccurs	
Date	TemporalRelation	refers	Date and Quantity are a special case because these entities are considered as abstract entities in SUMO. Thus their relation with other physical entities is established by the “refers” slot.
Quantity	Quantity	refers	

On the other hand context is mapped as the SUMO concepts “noun” and “verb” in accordance with the information gathered from the SL-tagger (refer to section 3). According to [1] the study of context’s effect in QA is one of the complex issues that requires formal models as well as experimentation in order to improve the performance of QA systems. The context considered for our preliminary experiments consists of the four verbs or nouns both at the left and right of its corresponding NE. Despite the fact that this parameter was chosen empirically, the results over the test collection are encouraging (refer to section 4.2).

Table 2 shows a subset of the instances collected from a sample document. Each row corresponds to an instance, and each concept is in bold font.

Table 2. An extract of SUMO instances gathered from a sample document.

```
<sumo:FactualText rdf:about="#010698-1Lunes">
  <sumo:refers rdf:resource="#Cárdenas"/>
  <sumo:refers rdf:resource="#PNR"/>
  <sumo:refers rdf:resource="#Tamaulipas"/>
  <sumo:refers rdf:resource="#1931"/>
</sumo:FactualText>
```

```
<sumo:Human rdf:about="#Cárdenas">
  <sumo:cooccur rdf:resource="#presidente"/>
  <sumo:cooccur rdf:resource="#PNR"/>
  <sumo:cooccur rdf:resource="#echar"/>
  <sumo:cooccur rdf:resource="#mano"/>
</sumo:Human>
```

```
<sumo:Organization rdf:about="#PNR">
  <sumo:cooccur rdf:resource="#presidente"/>
  <sumo:cooccur rdf:resource="#echar"/>
  <sumo:cooccur rdf:resource="#mano"/>
  <sumo:cooccur rdf:resource="#Ersatz"/>
  <sumo:cooccur rdf:resource="#democracia"/>
</sumo:Organization>
```

```
<sumo:GeographicArea rdf:about=”#Tamaulipas”>
  <sumo:cooccur rdf:resource="#gobierno"/>
  <sumo:cooccur rdf:resource="#subir"/>
  <sumo:cooccur rdf:resource="#partido"/>
  <sumo:cooccur rdf:resource="#Partido_Social_Fronterizo"/>
</sumo: GeographicArea>
```

```
<sumo:TemporalRelation rdf:about="#1931">
  <sumo:refers rdf:resource="#echar"/>
  <sumo:refers rdf:resource="#mano"/>
  <sumo:refers rdf:resource="#Ersatz"/>
  <sumo:refers rdf:resource="#democracia"/>
  <sumo:refers rdf:resource="#vez"/>
  <sumo:refers rdf:resource="#selección/>
  <sumo:refers rdf:resource="#candidato"/>
  <sumo:refers rdf:resource="#gobernador"/>
</sumo: TemporalRelation >
<sumo:Verb rdf:about="#echar"></sumo:Verb>
<sumo:Verb rdf:about="#subir"></sumo:Verb>
<sumo:Noun rdf:about="#presidente"></sumo:Noun>
<sumo:Noun rdf:about="#gobierno"></sumo:Noun>
<sumo:Noun rdf:about="#partido"></sumo:Noun>
<sumo:Noun rdf:about="# Partido_Social_Fronterizo "></sumo:Noun>
<sumo:Noun rdf:about="#mano"></sumo:Noun>
<sumo:Noun rdf:about="#Ersatz"></sumo:Noun>
<sumo:Noun rdf:about="#democracia"></sumo:Noun>
<sumo:Noun rdf:about="#vez"></sumo:Noun>
<sumo:Noun rdf:about="#selección"></sumo:Noun>
<sumo:Noun rdf:about="#candidato"></sumo:Noun>
<sumo:Noun rdf:about="#gobernador"></sumo:Noun>
```

3 Extraction Process

We describe in this section the NE tagger used in order to extract the entities and their contexts that will be used to represent the documents. This NE tagger is also used to extract NEs in the questions which will help us exploit the representation model for question resolution. As mentioned earlier, this extraction process is performed off-line. Once we have extracted the entities and their contexts, these are taken as instances of an upper level ontology as described in section 2.2.

A NE is a word or sequence of words that falls in one of these five categories: name of persons, organizations, locations, dates and quantities. There has been a considerable amount of work aimed to develop NE taggers with human-level performance. However, this is a difficult goal to achieve due to a common problem of all natural language processing tasks: ambiguity; another inconvenience is that documents are not uniform, their writing style, as well as their vocabulary change dramatically from one collection to another.

The NE tagger used in this work is that proposed by [11]. This system is based on training a Support Vector Machine (SVM) [10,12,13] classifier using as features the outputs of a handcrafted system together with information acquired automatically from the document, such as Part-of-Speech (POS) tags and capitalization information. The goal of this method is to reduce the effort in adapting a handcrafted NE extractor to a new domain. Instead of redesigning the grammars or regular expressions, and revising the lists of trigger words and gazetteers, we need only to build a training set by correcting, when needed, the outputs of the handcrafted system.

The starting handcrafted system used is considered by Solorio and López (SL) tagger as a black box, in particular, the system developed by [2] was used. The system classifies the words in the documents into the following six categories: Persons, Organizations, Locations, Dates, Numeric Expressions, and “none of the above”. Then each word in the documents has as features the output of the handcrafted system, their POS tag and their capitalization information (first letter capitalized, all letters capitalized, etc.). A previously trained SVM assigns the final NE tags using the features mentioned above. This process can be considered as a stacking classifier, in the first stage a handcrafted system assigns NE tags to the document, and then these tags (corrected if necessary) are used as inputs to the SVM classifier which decides the final NE tags.

In order to show an example of how this NE tagger performs, we present here a comparison between the handcrafted system and the tagger from Solorio and López. Table 3 shows the results of tagging questions that can be answered using the model proposed here. In this table, we only show the named entities from the questions. As it can be seen, the SL tagger improves the accuracy of the handcrafted system. In this example, the SL tagger corrects 6 tags that were originally misclassified by the handcrafted system.

Table 3. Comparison between the handcrafted system (HS) and that of Solorio and López (SL). Cases where the HS tagger misclassifies NEs that are correctly classified by the SL tagger are in bold. The asterix (*) marks cases where the SL tagger misclassifies a NE correctly classified by the HS tagger.

Named Entity	HS tags	SL tags	True NE tag
Unión_de_Cineastas_de_Rusia	Organization	Organization	Organization
Director_de_Aeroméxico*	Person	Organization	Person
Irán	**Organization**	**Location**	**Location**
Copa_Mundial_de_Fútbol	**Person**	**Organization**	**Organization**
Irán	**Organization**	**Location**	**Location**
Irán-Estados_Unidos*	Location	Organization	Location
Aeroméxico	**Person**	**Organization**	**Organization**
Aeroméxico	**Person**	**Organization**	**Organization**
Ruanda	Location	Location	Location
Mundial_Francia	Organization	Organization	Organization
Consejo_de_Ministros_de_Líbano	Organization	Organization	Organization
OTAN	Organization	Organization	Organization
Estados_Unidos	**Organization**	**Location**	**Location**
Accuracy	**53%**	**84%**	

4 Case of Study

This section presents a schema for the application of the proposed model to an experimental –and yet simple– QA system. In this case the searching process uses only the information considered by the model.

The algorithm shows the appropriateness of the representation in searching for answers to factual questions. The following subsection describes the general algorithm and its application over a sample collection of news in Spanish. Given the limitation of space no implementation details are given.

4.1 The Algorithm

The algorithm is based in two key assumptions:

First, the kind of the question defines the class of NE to search. Generally speaking, factual questions do not rely on the predicate of the sentence, but on the subject, the characteristics of the question, or on some other sentence element. In this way, by the interrogative adverb (Wh-word) employed in the question, it is possible to infer the role of the NE required as an answer. For instance, *"¿Quién es el presidente de México?"*[6] requires to be answered with a NE of the class person (human). Of

[6] Who is the president of México?

course not all interrogative adverbs define the kind of NE for the answer, e.g. *"¿Cuál es el nombre del presidente de México?"*[7]. For now, the algorithm is focused on partial interrogative questions whose answer role could be immediately identified by the interrogative adverb employed.

Second, from the question itself two kinds of information can be extracted: its NEs and the lexical context of the question. With the proposed model, all the NEs mentioned in a given document can be known beforehand. Thus the NEs from the question become key elements in order to define the document set more likely to provide the answer. For instance, in any of the sample questions above, the NE "Mexico" narrows the set of documents to only those containing such NE. At the same time, another assumption is that the context in the neighborhood of the answer has to be similar to the lexical context of the question. Once more, from the sample question, the fragment "even before his inauguration as president of Mexico, Vicente Fox…" contains a lexical context next to the answer which is similar to the question.

Following is the algorithm in detail:

1. Identify the type of NE-answer for a given question. We are limited by the set of NEs in the model (persons, organizations, locality, date & time, and quantity).
2. Extract NEs contained in the question and starting from them identify the appropriate document subset.
3. Retrieve all candidate NEs and their local lexical context (as detailed by the model) starting from those identified in step 2.
4. Compute the similarity between question context and those of the candidate NEs.
5. Rank the candidate NE in decreasing order of similarity.
6. Report the top NEs as possible answers

4.2 Results

This subsection shows the application of the algorithm just described on a small text collection of news in Spanish. The collection News94 consists of a set of 94 news (see table 4 for collection details). These documents contain national and international news from the years 1998 to 2000. Regarding extracted information, the total of NEs obtained from this collection was 3191 (table 4 also shows totals by class).

Table 4. Main data of collection News94

Collection	**Size**	**Number of documents**	**Average document size**	**Number of pages**	**Number of lexical forms**	**Number of terms**
	372 Kb	94	3.44 Kb	124	11,562	29,611
Entities	**Date & Time**	**Locality**	**Organization**	**Person**	**Quantity**	**Others**
	266	570	1094	973	155	133

[7] What is the name of the president of Mexico?

The processing of the question: *“¿Quién era el presidente del PNR en 1931?”*[8] was done as follows:

1. Identify the class of the NE to search. Given that the interrogative adverb is *“quién” (who)*, the class is person (human).
2. Extract the NEs in the question. These are: PNR (Organization), present in the set of documents {0,13,86}; 1931 (Date), found in the set of documents {0}. As a consequence, the subset of documents is {0}.
3. Retrieve all NEs of class person (human) from the document ‘0’
4. Compute the similarity between the question context, that is {ser (be), presidente (president), PNR, 1931} and candidate NEs contexts. Table 5 shows the computed similarity.

5 y 6. {Cárdenas, Cárdenas_Presidente}

Table 5. Candidate NEs, their context and similarity

Context	**NE**	**Sim.**
{creador, 30, año, plebiscito, añoranza, embellecer, muerte}	Portes_Gil	0
{presidente, PNR, echar, mano}	Cárdenas	2
{prm, fundar, carácter, otorgar, ser, forma, permanecer, candidatura}	Cárdenas_Presidente	1
{arribar, 1964, pri, presidencia, Carlos, juventud, líder, camisa}	Madrazo	0
{Madrazo, arribar, 1964, pri, juventud, líder, traer}	Carlos	0
{pri, faltar, mano, gato}	Madrazo	0
{zurrar, Sinaloa, Madrazo, corto}	Polo_Sánchez_Celis	0
{Sinaloa, zurrar, Polo_Sánchez_Celis, corto, 11, mes, salir}	Madrazo	0
{pierna, cola, pri, salir, diputado, pelea, deber}	Polo_Martínez_Domínguez	0

In this example, “Cárdenas” is the correct answer, and the original text passage is shown in table 6.

Table 6. Passage with the answer to the sample question

“**Cárdenas** como presidente del PNR echó mano del Ersatz de democracia en 1931 por vez primera en la selección de candidatos a gobernadores”[9].

Table 7 shows a subset of the questions used in our preliminary experiments. A total of 30 questions were proposed by 5 assessors for this experiment. From these questions only 22 were classified as factoid and were evaluated. Results show that for 55% of the questions, the answer is found in the first NE, and that 82% of the questions are correctly responded within the top-5 NEs.

Despite of the informal evaluation of the algorithm and the small size of the collection, we found these results very encouraging, hinting the appropriateness of the proposed model and the likely robustness of the QA algorithm.

[8] Who was the president of the PNR in 1931?

[9] **Cardenas**, as president of the PNR made use of the Ersatz of democracy in 1931, for the first time in the selection of candidates for governor.

Table 7. Subset of testing questions.

Question	1st Answer	Correct Answer
¿Quién es el presidente de la Unión de Cineastas de Rusia? (Who is the president of the Moviemakers Union of Rusia?)	El Barbero de Siberia	Nikita Mijalkov
¿Quién ha impulsado el desmantelamiento del presidencialismo? (Who encouraged the dismantling of presidentialism?)	Zedillo	Zedillo
¿Cuándo calificó por última vez Irán para una Copa Mundial de Futbol? (When was the last time Iran classified for a Soccer World Cup?)	1978	1978
¿Quién es el presidente de Irán? (Who is president of Iran?)	Muhamad Khatami	Muhamad Khatami
¿Quién es la dirigente del sindicato de sobrecargos de Aeroméxico? (Who is the union leader of flight attendants of Aeromexico?)	Carlos_Ruíz_Sacristán	Alejandra Barrales Magdaleno
¿Cuántas personas fueron asesinadas en Ruanda durante 1994? (How many people were murdered in Rwanda during 1994?)	500,000	Más de 500 mil
¿Cuántos jugadores de futbol participarán en el Mundial de Futbol Francia 1998? (How many soccer player will participate in the World Soccer Cup of France 1998?	------	704
¿Cuándo aprobó el senado la ampliación de la OTAN? (When did the senate approved the expansion of OTAN?)	30 de abril	30 de abril de 1998
Correct answer in the first NE	**55%**	
Correct answer within the top-5 NE	**82%**	

5 Conclusions

The proposed model can be an initial step toward document representation for specific tasks, such as QA as detailed. This representation is functional because captures valuable information that allows performing retrieval and extraction processes for QA in an easier and more practical way. Some important features of this model are that it considers a broader classification of NEs which improves the precision of the system; it also accelerates the whole process by searching only in the corresponding named entity class that is believed to contain the answer, instead of searching the answer in the whole document.

Besides, the representation is expressed in a standardized language as DAML+OIL –soon could be OWL–, in the direction of the next Web generation. This could yield to the exploitation of this document representation in multilingual settings for QA either in stand alone collections or the Semantic Web.

Preliminary results exploiting the representation of documents as proposed by the model were very encouraging. The context similarity assessment method has to be refined and additional information can be taken into account, e.g. proximity. We are also in the process of experimenting with large text collections and questions sets supplied by international conferences on Questions Answering systems such as TREC or CLEF. In further developments of this model we pretend to refine the classification of named entities in order to take full advantage of the ontology.

Acknowledgements. This work was done under partial support of CONACYT (Project Grant U39957-Y), SNI-Mexico, and the Human Language Technologies Laboratory of INAOE.

References

1. Burger, J. et al. *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A).* NIST 2001.
2. Carreras, X. and Padró, L. *A Flexible Distributed Architecture for Natural Language Analyzers*. In Proceedings of the LREC'02, Las Palmas de Gran Canaria, Spain, 2002.
3. Cowie J., et al., *Automatic Question Answering*, Proceedings of the International Conference on Multimedia Information Retrieval (RIAO 2000)., 2000.
4. Hirshman L. and Gaizauskas R. *Natural Language Question Answering: The View from Here*, Natural Language Engineering 7, 2001.
5. Magnini B., Romagnoli S., Vallin A., Herrera J., Peñas A., Peinado V., Verdejo F. and Rijke M. *The Multiple Language Question Answering Track at CLEF 2003*. CLEF 2003 Workshop, Springer-Verlag.
6. Mann, G.S. *Fine-Grained Proper Noun Ontologies for Question Answering*, SemaNet'02: Building and Using Semantic Networks, 2002.
7. Niles, I. and Pease A., *Toward a Standard Upper Ontology*, in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), 2001.
8. Prager J., Radev D., Brown E., Coden A. and Samn V. *The Use of Predictive Annotation for Question Answering in TREC8*. NIST 1999.
9. Ravichandran D. and Hovy E. *Learning Surface Text Patterns for a Question Answering System*. In ACL Conference, 2002.
10. Schölkopf, B. and Smola A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2001.
11. Solorio, T. and López López A. *Learning Named Entity Classifiers using Support Vector Machines*, CICLing 2004, LNCS Springer-Verlag, Feb. 2004, (to appear).
12. Stitson, M.O., Wetson J.A.E., Gammerman A., Vovk V., and Vapnik V. *Theory of Support Vector Machines*. Technical Report CSD-TR-96-17, Royal Holloway University of London, England, December 1996.
13. Vapnik, V. *The Nature of Statistical Learning Theory*, Springer, 1995.
14. Vicedo, J.L., Izquierdo R., Llopis F. and Muñoz R., *Question Answering in Spanish*. CLEF 2003 Workshop, Springer-Verlag.
15. 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.