Using Machine Learning and Text Mining in Question Answering

Antonio Juárez-González, Alberto Téllez-Valero, Claudia Delicia-Carral,
Manuel Montes-y-Gómez and Luis Villaseñor-Pineda

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{antjug,albertotellezv,cdenicia,mmontesg,villasen}@inaoep.mx

Abstract. This paper describes a QA system centered in a full data-driven architecture. It applies machine learning and text mining techniques to identify the most probable answers to factoid and definition questions respectively. Its major quality is that it mainly relies on the use of lexical information and avoids applying any complex language processing resources such as named entity classifiers, parsers and ontologies. Experimental results on the Spanish Question Answering task at CLEF 2006 show that the proposed architecture can be a practical solution for monolingual question answering by reaching a precision as high as 51%.

1 Introduction

Current information requirements suggest the need for efficient mechanisms capable of interacting with users in a more natural way. Question Answering (QA) systems have been proposed as a feasible option for the creation of such mechanisms [1]. Recent developments in QA use a variety of linguistic resources to help in understanding the questions and the documents. The most common linguistic resources are: part-of-speech taggers, parsers, named entity extractors, dictionaries, and WordNet [2, 3, 4, 5, 10]. Despite the promising results of these approaches, they have two main inconveniences. On the one hand, the construction of such linguistic resources is a very complex task. On the other hand, their performance rates are usually not optimal.

In this paper we present a QA system that can answer factoid and definition questions. This system is based on a full data-driven approach that requires a minimum knowledge about the lexicon and the syntax of the specified language. It is built on the idea that the questions and their answers are commonly expressed using the same set of words. Therefore, it simply uses lexical information to identify the relevant document passages and to extract the candidate answers.

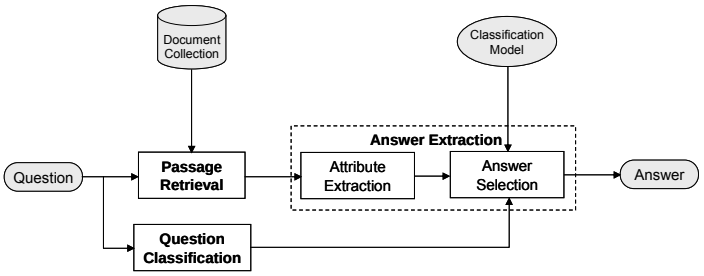
The proposed system continues our previous work [6] by considering a lexical data-driven approach. However, it presents two important modifications. First, it applies a supervised approach instead of a statistical method for answering factoid

questions. Second, it answers definition questions by applying lexical patterns that were automatically constructed rather manually defined.

The following sections give some details of the system. In particular, section 2 describes the method for answering factoid questions, section 3 explains the method for answering definition questions, and section 4 discusses the results achieved by our system at the CLEF 2006 Spanish Question Answering task.

2 Answering Factoid Questions

Figure 1 shows the general process for answering factoid questions. It considers three main modules: *passage retrieval*, where the passages with a high probability of containing the answer are recovered from the document collection; *question classification*, where the type of expected answer is determined; and *answer extraction*, where candidate answers are selected using a machine-learning approach, and the final answer recommendation of the system is produced. The following sections describe each of these modules.



Fig, 1. Process for answering factoid questions

2.1 Passage Retrieval

The passage retrieval (PR) method is especially suited to the QA task. It allows passages with the highest probability of containing the answer to be retrieved, 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 highest weights.

Details about the PR method can be found in [7].

2.2 Question Classification

Given a question, this module is responsible for determining the semantic class of the expected answer. This information will be used later to reduce the searching space. The idea is to focus the answer extraction only on those text fragments related to the expected type of answer.

Our system prototype implements this module following a direct approach based on regular expressions. It only considers three general semantic classes of expected answers: dates, quantities and names (i.e., general proper nouns).

2.3 Answer Extraction

Answer extraction aims to establish the best answer for a given question. It is based on a supervised machine-learning approach. It consists of two main modules, one for attribute extraction and the other for answer selection.

Attribute extraction. First, the recovered passages are processed in order to identify all text fragments related to the expected type of answer. Each one of the identified text fragments is considered as a "candidate answer". It is important to mention that this analysis is also based on a set of regular expressions.

In a second step, the lexical context of each candidate answer is analyzed with the aim of constructing its formal representation. In particular, each candidate answer is represented by a set of 17 attributes, clustered in the following groups:

1. Attributes that describe the complexity of the question, for instance, its length (number of non-stopwords).
2. Attributes that measure the similarity between the context of the candidate answer and the given question. Some of them describe the word overlap between the question and the context of the candidate answer. Some others measure the density of the question words in the context of the candidate answer.
3. Attributes that indicate the relevance of the candidate answer in reference to the set of recovered passages. Some of these attributes are: the redundancy of the candidate answer in the set of recovered passages, and the position of the passage containing the candidate answer.

Answer Selection. This module is based on a machine-learning approach. Its purpose is to select, from the set of candidate answers, the one with the maximum probability of being the correct answer. In particular, this module is implemented by a Naïve Bayes classifier, which was constructed using as training set the questions and documents from the previous CLEF campaigns.

3 Answering Definition Questions

Figure 2 shows the general scheme of our method for answering definition questions. It consists of three main modules: a module for the discovery of definition patterns, a module for the construction of a general definition catalogue, and a module for the extraction of the candidate answer. The following sections describe in detail these modules.

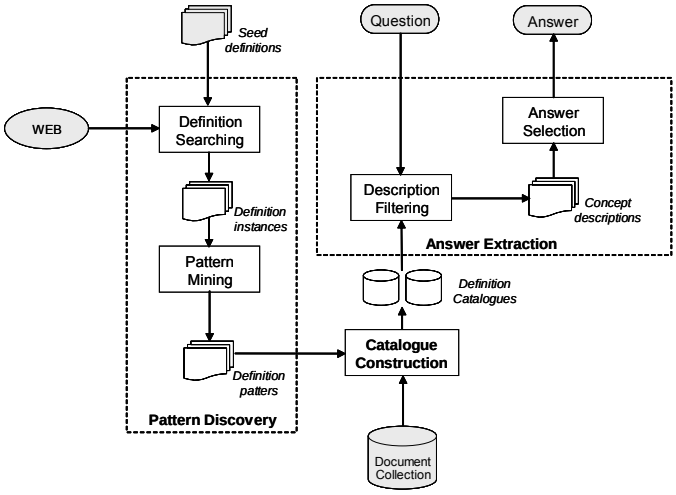


Fig. 2. Process for answering definition questions

It is important to mention that this method is especially suited to answering definition questions as defined in CLEF. That is, questions asking for the position of a person, e.g., Who is Vicente Fox?, and for the description of concept, e.g., What is CERN? or What is Linux?.

It is also important to notice that the processes for pattern discovery and catalogue construction are done offline, while the answer extraction is done online, and that in contrast to traditional QA approaches, the proposed method does not use any module for document or passage retrieval.

3.1 Pattern Discovery

The module for pattern discovery uses a small set of concept-description pairs to collect from the Web an extended set of definition instances. Then, it applies a text mining method to the collected instances to discover a set of definition surface patterns. The idea is to capture the definition conventions through their repetition. This module involves two main subtasks:

Definition searching. This task is triggered by a small set of empirically defined concept-description pairs. The pairs are used to retrieve a number of usage examples from the Web[1]. Each usage example represents a definition instance. To be relevant, a definition instance must contain the concept and its description in one single phrase.

Pattern mining. It is divided into three main steps: data preparation, data mining and pattern filtering. The purpose of the data preparation phase is to normalize the input data. It transforms all definition instances into the same format using special tags for the concepts and their descriptions. It also indicates with a special tag the concepts expressing proper names.

In the data mining phase, a sequence mining algorithm [9] is used to obtain all maximal frequent sequences of words[2], punctuation marks and tags from the set of definition instances. The sequences express lexicographic patterns highly related to concept definitions.

Finally, the pattern-filtering phase allows choosing the more discriminative patterns. It selects the patterns satisfying the following general regular expressions:

```
        <left-string> DESCRIPTION <middle-string> CONCEPT <right-string>
        <left-string> CONCEPT <middle-string> DESCRIPTION <right-string>
  <left-string> DESCRIPTION <middle-string> PROPER_NAME_CONCEPT <right-string>
  <left-string> PROPER_NAME_CONCEPT <middle-string> DESCRIPTION <right-string>
          <left-string> DESCRIPTION <middle-string> PROPER_NAME_CONCEPT
         PROPER_NAME_CONCEPT <middle-string> DESCRIPTION <right-string>
                  <left-string> DESCRIPTION PROPER_NAME_CONCEPT
                 PROPER_NAME_CONCEPT DESCRIPTION <right-string>
```

The idea of the pattern discovery process is to obtain several surface definition patterns, starting with a small set of concept-description example pairs (details about this process can be found in [8]). For instance, using description seeds "*Wolfgang Clement – German Federal Minister of Economics and Labor*" and "*Vicente Fox – President of Mexico*", at the end we could obtain definition patterns such as "*, the <DESCRIPTION>, <CONCEPT>, says*" and "*the <DESCRIPTION> <PROPER_NAME_CONCEPT>*". It is important to notice that the discovered patterns may include words and punctuation marks as well as proper name tags as frontier elements.

3.2 Catalogue Construction

In this module, the discovered definition patterns are applied over the target document collection. The result is a set of matched text segments that presumably contain a concept and its description. The definition catalogue is created gathering all matched segments.

---

[1] At present we are using Google for searching the Web.

[2] The word sequence p is frequent in the collection D if it occurs in at least σ texts of D, and it is maximal if there does not exist any sequence p' in D such that p is a subsequence of p' and p' is also frequent in D.

3.3 Answer Extraction

This module handles the extraction of the answer for a given definition question. Its purpose is to find the most adequate description for a requested concept from the definition catalogue. The definition catalogue may contain a huge diversity of information, including incomplete and incorrect descriptions for many concepts. However, it is expected the correct information will be more abundant than incorrect information. This expectation supports the idea of using a frequency criterion and a text mining technique to distinguish between the adequate and the improbable answers to a given question. This module considers the following steps:

Description filtering. Given a specific question, this procedure extracts from the definition catalogue all descriptions corresponding to the requested concept. As we mentioned, these "presumable" descriptions may include incomplete and incorrect information. However, it is expected that many of them will contain, maybe as a substring, the required answer.

Answer selection. This process aims to detect a single answer to the given question from the set of extracted descriptions. It is divided into two main phases: data preparation and data mining.

The data preparation phase focuses on homogenizing the descriptions related to the requested concept. The main action is to convert these descriptions to a lower case format. In the data mining phase, a sequence mining algorithm [9] is used to obtain all maximal frequent word sequences from the set of descriptions. Finally, each sequence is ranked based on the frequency of occurrence of its subsequences in the whole description set [8]. The best one is given as the final answer.

Figure 3 shows the process of answer extraction for the question "*Who is Diego Armando Maradona?*". First, we obtained all descriptions associated with the requested concept. It is clear that there are erroneous or incomplete descriptions (e.g. "*Argentina soccer team*"). However, most of them contain a partially satisfactory explanation of the concept. Actually, we detected correct descriptions such as "*captain of the Argentine soccer team*" and "*Argentine star*". Then, a mining process allowed detecting a set of maximal frequent sequences. Each sequence was considered a candidate answer. In this case, we detected three sequences: "*argentine*", "*captain of the Argentine soccer team*" and "*supposed overuse of Ephedrine by the star of the Argentine team*". Finally, each candidate answer was ranked based on the frequency of occurrence of its subsequences in the whole description set. In this way, we took advantage of the incomplete descriptions of the concept. The selected answer was "*captain of the Argentine national football soccer team*", since it was formed from frequent subsequences such as "*captain of the*", "*soccer team*" and "*argentine*".

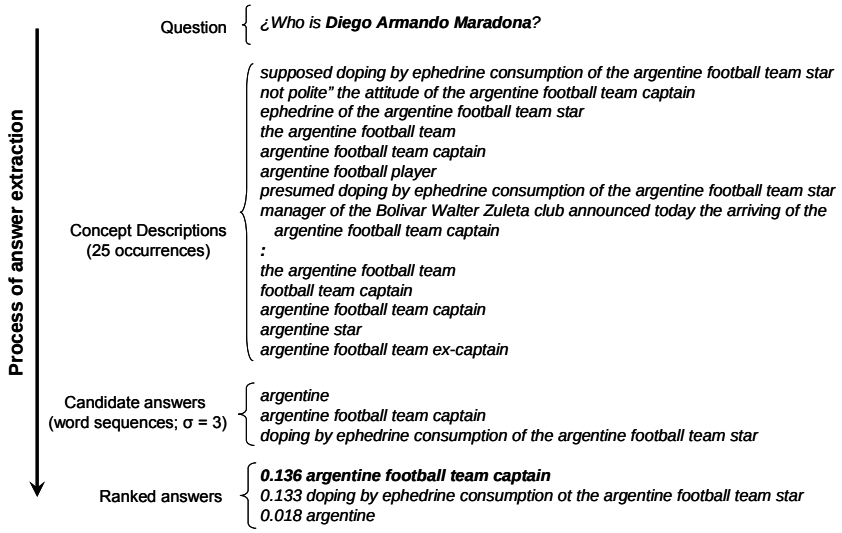


Fig. 3. Data flow in the answer extraction process

4 Evaluation Results

This section describes the experimental results related to our participation at the QA@CLEF2006 monolingual track for Spanish. It is important to remember that this year the question type (e.g., factoid, definition, temporal or list) was not included as a data field in the question test file. Therefore, each participant had to automatically determine the kind of question.

Our system prototype, as described in the previous sections, can only deal with factoid and definition questions. From the 200 test questions, it treats 144 as factoid questions and 56 as definition questions. Table 1 details our results on answering both types of questions considering the evaluation given by CLEF judges and our own evaluation.

Table 1. Evaluation of the system on answering factoid and definition questions

	Accuracy	**R**	**W**	**X**	**U**	**Factoid Questions (144)**	**Definition Questions (56)**
CLEF evaluation	51%	102	86	3	9	59 (40.9%)	43 (76.7%)
Local evaluation	53%	106	86	3	5	63 (43.75%)	43 (76.7%)

CLEF evaluation gave our system an accuracy of 51%, by answering 102 questions correctly, 59 factoids and 43 definitions. In contrast, in our local evaluation we obtain an accuracy of 53%. This difference was caused by a discrepancy in the

evaluation of four factoid questions about dates. The answers for these questions lacked a year string in the passages, therefore we decided to extract the year string from the document ID. This way, we obtained an entire date answer (consisting of a day, month and a year) completely supported considering both the passage and the document ID. Nevertheless, CLEF judges considered these four questions as unsupported, affecting the overall accuracy of our system. As an example, table 2 shows our answer for one of these questions.

Table 2. An adapted answer for a date question

Document ID	Question	Answer	Passage
EFE**19950608**-05247	When was the G7 meeting in Halifax?	June 15th to 17th **1995**	…the meeting of the seven most industrialized countries (G7) in Halifax (Canada), from **June 15th to 17th**, and the European Council of Cannes- figure in the order of the day of this informal summit.

Lastly, it is important to point out some facts about our participation at the QA@CLEF 2006 [11]. First, our prototype was one of the four systems that went beyond the barrier of 50% in accuracy. Second, it was the best system for answering definition questions. Third, the overall evaluation accuracy of this year exercise (51%) was 10-points over our result for last year [8].

5 Conclusions and Future Work

This paper presented a question answering system that can answer factoid and definition questions. This system is based on a lexical data-driven approach. Its main idea is that the questions and their answers are commonly expressed using almost the same set of words, and therefore, it simply uses lexical information to identify the relevant passages as well as the candidate answers.

The answer extraction for factoid questions is based on a machine learning method. Each candidate answer (an uppercase word denoting a proper name, a date or a quantity) is represented by a set of lexical attributes and a classifier determines the most probable answer for the given question. The method achieved good results, however it has two significant disadvantages: (*i*) it requires a lot of training data, and (*ii*) the detection of the candidate answers is not always (not for all cases, nor for all languages) an easy task to perform with high precision.

On the other hand, the answer extraction for definition questions is based on a text mining approach. The proposed method uses a text mining technique (namely, a sequence mining algorithm) to discover a set of definition patterns from the Web as well as to determine with finer precision the answer to a given question. The achieved results were especially good, and they showed that a non-standard QA approach, which does not contemplate an IR phase, can be a good scheme for answering definition questions.

As future work we plan to consider syntactic attributes in the process of answering factoid questions. These attributes will capture the matching of syntactic trees as well as the presence of synonyms or other kinds of relations among words. In addition, we are also planning to improve the final answer selection by applying an answer validation method.

Acknowledgments. This work was done under partial support of CONACYT (Project Grant: 43990). We would also like to thank the CLEF organizing committee as well as to the EFE agency for the resources provided.

References

1. Magnini B., Vallin A., Ayache C., Erbach G., Peñas A., de Rijke M., Rocha P., Simov K. and Sutcliffe R., *Overview of the CLEF 2004 Multilingual Question Answerig Track*. In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2004), Bath, UK, September 2004.
2. Ferrández S., López-Moreno P., Roger S., Ferrández A., Peral J., Alvarado X., Noguera E., and Llopis F. *AliQAn and BRILI QA Systems at CLEF 2006*. In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), September 2006.
3. Buscaldi D., Gomez J.M., Rosso P., and Sanchis E. *The UPV at QA@CLEF 2006*. In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), September 2006.
4. de-Pablo-Sánchez C., González-Ledesma A., Moreno A., Martínez-Fernández J.L., and Martínez P. *MIRACLE at the Spanish CLEF@QA 2006 Track*. In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), September 2006.
5. Ferrés D. Kanaan S., González E., Ageno Al, Rodríguez H. and Turmo J., *The TALP-QA System for Spanish at CLEF-2005*. In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2005), Vienna, Austria, September 2005.
6. Montes-y-Gómez M., Villaseñor-Pineda L., Pérez-Coutiño M., Gómez-Soriano J.M., Sanchis-Arnal E. and Rosso P., *INAOE-UPV Joint Participation at CLEF 2005: Experiments in Monolingual Question Answering*. In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2005), Vienna, Austria, September 2005.
7. Gómez-Soriano J.M., Montes-y-Gómez M., Sanchis-Arnal E., Villaseñor-Pineda L. and Rosso P., *Language Independent Passage Retrieval for Question Answering*. In Proceedings for the Fourth Mexican International Conference on Artificial Intelligence (MICAI 2005), Monterrey, Nuevo León, México, November 2005.
8. Denicia-Carral C., Montes-y-Gómez M., Villaseñor-Pineda L. and García-Hernández R., *A Text Mining Approach for Definition Question Answering*. In Proceedings for the Fifth International Conference on Natural Language Processing (FinTal 2006), Turku, Finland, August 2006.

9. García-Hernández R., Martínez-Trinidad F. and Carrasco-Ochoa A., *A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection*. In Proceedings for the Seventh International Conference on Computational Linguistics and text Processing (CICLing 2006), Mexico City, Mexico, February 2006.
10. Cassan A., Figueira H., Martins A., Mendes A., Mendes P., Pinto P., and Vidal D. *Priberam's question answering system in a cross-language environment*. In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), September 2006.
11. Magnini B., Giampiccolo D., Forner P., Ayache C., Jijkoun V., Osenova P., Peñas A., Rocha P., Sacaleanu B., and Sutcliffe R. *Overview of the CLEF 2006 Multilingual Question Answering Track.* In Working notes for the Cross Language Evaluation Forum Workshop (CLEF 2006), September 2006.