QA UdG-UPC System at TREC-12

Marc Massot
Dept. Informàtica i Matemàtica Aplicada
Universitat de Girona
marc@ima.udg.es

Horacio Rodríguez
TALP Researh Center
Universitat Politècnica de Catalunya
dferres@lsi.upc.es

Daniel Ferrés
TALP Researh Center
Universitat Politècnica de Catalunya
dferres@lsi.upc.es

Abstract

This paper describes a prototype multilingual Q&A system that we have designed to participate in the Q&A Track of TREC-12. The system answer concrete responses, then we participate in the Q&A main task for factoid questions. The main areas of our system are: (1) Inductive Logic Programming to learn the question type, (2) Clustering of Named Entities to improve Information Retrieval and (3) Semantic relations and EuroWordNet synsets to perform a language-independent answer extraction.

1. Introduction

This paper describes a prototype Q&A system we have designed to participate in the Q&A Track of TREC-12. Our aim has been to build a system as much as possible language independent, where language dependent modules could be changed for allowing the system to be applied to different languages. In this way we have developed in parallel two different Q&A systems, one for English and another for Spanish.

As our research group has mainly focused in building resources and tools for NLP in Spanish, we have directly applied these tools and resources in our system. For English system we have used, when possible, publicly available resources or adapted our own tools.

In this paper we present the overall architecture of the system, we describe briefly its main parts, focusing on the language independent ones, and we present a preliminary evaluation of the prototype presented at the TREC-12 competition.

Our system was designed to participate in the Q&A main task for factoid questions. Thus, we develop a system to answer questions with a concrete response. We structure the remaining part of the paper as follows. In Section 2, we first give an overview of the system

and then focusing on every subsystem. Finally, in section 3, we evaluate the results of this participation and we detail some conclusions.

2. The System

Overview

The system architecture follows the most commonly used schema, splitting the process into three phases that are performed in turn. Several iterations can be carried out within these phases to achieve their goals but once one phase is finished there is no possibility to return to previous phases. There are three main subsystems, one corresponding to each phase:

1. Question processing (QP)
2. Passage retrieval (PR)
3. Answer extraction (AE)

These subsystems are described below. Some pre-processing has been done on the document collection (the Acquaint corpus in this case). We will describe this issue when we present the PR subsystem.

Language dependent components are included only within the QP and AE subsystems.

Question Processing

The main goal of this subsystem is to classify the question regarding the kind of expected answer and to attach the information needed for the following subsystems. For PR the information needed is basically lexical (lists of keywords) and for AE lexical, syntactic and semantic. We have tried to represent all these kind of information using a language independent formalism. In particular we use the same semantic primitives and relations for the two languages (English and Spanish) involved in our system.

This subsystem uses a great amount of linguistic resources for performing its task. As our goal is processing questions in Spanish and English, both with independent linguistic resources and tools, we need mapping tools for providing information for the following subsystems in an uniform representation.

The tools used for the Spanish version are those of the group of NLP of the UPC (see [Atseries et al, 1998] for a description of these tools). The question is analyzed with a pipe including the following processors:

- **ms-analyze**, that performs tokenization, morphological analysis (including identification of quantities, dates, multiword terms, etc.), and POS tagging. As a result we obtain a sequence of tokens with POS and lemma.
- **tacat**, a partial parser that obtains nominal, prepositional and verbal phrases.

- **NERC**, a Named Entity Recognizer and Classifier that identifies the NE occurring in the question and classifies them in basic categories (person, place, organization,…). See [Carreras et al, 2002]
- Finally we obtain and attach semantic information using EWN[1]. The information obtained and used for further processing consists of the list of synsets (with no attempt to Word Sense Disambiguation), the list of hyperonyms of each synset (up to the top of each hyperonymy chain), the EWN's Top Concept Ontology, TCO class [Rodriguez et al,1998],and the Magnini's Domain Code [Magnini, Cavaglià, 2000].

For English version, we have adapted some of our tools or used publicly available ones for getting the same information using the same representation formalism.

- **TnT** [Brants, 2000], for the morphologic information, As TnT does not provide the lemma we have used the lemmatizer included in Princeton's WordNet software for covering this functionality.
- **MINIPAR** [Lin, 1998], to perform full parsing. A post-process has been needed for representing the output in a way compatible with tacat's output.
- The same NERC used for Spanish has been trained for English.
- A finer grained classifier for Geographic NE (those of type location) was developed, [Ferres et al, 2003a], devising a set of gazetteers with binary classifiers learned using an ILP learner, FOIL, [Quinlan, 1993].
- A gazetteer of acronyms obtained using a Decision Tree learning approach [Ferres et al, 2003b].
- A list of relations actor-action obtained through an analysis of the glosses of WordNet.

The result of applying these linguistic resources and tools, obviously language dependent, to the text of the question is represented in two structures:

- **Sint**, composed by two lists, one recording the information related to the syntactic components of the question (basically nominal, prepositional and verbal phrases) and the other collecting the information of dependencies and other relations between these components.
- **Sent**, that provides us with information for each lexical unit: the word form, the lemma, the POS, the semantic class (and subclass if available) of NE, the list of EWN synsets, the information associated to these synsets (TCO and DC) and, finally, whenever possible the verbs associated to the actor and the relations between locations and their gentile.

Once this information is obtained we are able to get the information relevant to the following tasks:

[1] http://www.illc.uva.nl/EuroWordNet/

- **Question type**. The most important information we need to extract from the question text is the Question Type, QT, because all the work the system has to perform for searching the answer is based on this issue. A failure on identifying QT practically disables the correct extraction of the answer. Currently we are working with about 50 QT. The QT tries to focus the type of the expected answer providing as well additional constraints on it. For instance, when the expected type of the answer is a person, two types of questions are considered, *Who_action*, which indicates that we are looking for a person who performs a certain action and *Who_person_quality*, that indicates that we are looking for a person having the desired quality. The action and the quality are the parameter of the corresponding QT. The following are examples of questions classified as *Who_action* type:

    - What is the name of the managing director of Apricot Computer?
    - Who won the Nobel Peace Prize in 1991?
    - Who is the writer of the book, "The Hobbit"?

  In order to determine the QT our system uses an Inductive Logic Programming (ILP) learner that learns, from a set of positive and negative examples, a set of weighted rules. We have used as learner the FOIL system [Quinlan, 1993]. We use FOIL for learning a different classifier (i.e. a set of rules) for each QT. As training set we have used the set of questions of TREC 8 and 9 (~900 questions) manually tagged and as test set the 500 questions of TREC 11. With these rules we have obtained an overall precision 68.54% and a recall of 60.00%. But the most of the errors are in similar QT categories, i.e. the 50% of errors for a generated when_begins are when_action questions, thus impact of these errors is minimum.
  For each classifier we have used as negative examples the questions belonging to the other classes. The features used for classifying are the following:

    - Word form
    - Word position in the question
    - Lemma
    - POS
    - Semantic class of NE, without subclassing
    - Synsets of the lemma
    - All hyperonyms for each synset without the information about the distance
    - TCO for each synset
    - Domain Codes for each synset
    - Main syntactic relations, subject and object relations

  The set of rules for each class has been manually revised and completed with a set of manually built rules (with lower weights) in order to assure a complete coverage. See below a couple of such rules:

    - A learned rule:

```
regla(non_human_actor_of_action,A,1) :-
  first_position(A,B),
  next_position(B,C),
```

```
    is_tco(cObject,C),
    is_domain(dTransport,C).
```

- The same rule after transformation(performed for the sake of efficiency):

```
regla(non_human_actor_of_action,A,1,[],TT) :-
  sent(A,_,TT), TT=[_,W2|_],
  has_tco(W2,cObject),
  has_domain(W2,dTransport).
```

- A manual rule:

```
regla(non_human_actor_of_action,A,994,[T1,T3],T) :-
  sent(A,_,[T1|T]),
  the_lema(T1,lema("which")),
  has_chunk_with_hyperonym(_,T,[T2|TT],
    [sArtifact,sObject,sAnimal],T3),
  the_pos(T2,pos("IN")),
  not(has_term_with_pos(TT,pos("JJS"),_)).
```

- **Environment.** Under this term we collect the semantic relations that hold between the different components identified in the question text. These relations are organized into an ontology of about 100 semantic classes and 25 relations (mostly binary) between them. Both classes and relations are related by taxonomic links. The ontology has been manually built and tries to reflect what is needed for an appropriate representation of the semantic environment of the question (and the expected answer). For instance, *Action* is a class and *Human_action* another class related to *Action* by an *is_a relation*. In the same way, *Human* is a subclass of *Entity*. *Actor_of_action* is a binary relation (between a *Human_action* and a *Human*). When a question is classified as *Who_action* an instance of the class *Human_action* has to be located in the question text. Later, in the AE phase, an instance of *Human_action* has to be located in the selected passages and an instance of *Human* related to it by the *Actor_of_action* relation has to be extracted as candidate to be the answer.

  The environment of the question is obtained from the syntactic information (**sint**) and the semantic information included in **sent**. For performing this task a set of about 150 rules has been manually built.

  For instance, for the question:

  Who won the Nobel Peace Prize in 1991?

  the following environment was obtained:

  action(A, won),
  time_of_event(A, T),
  year(T, 1991),
  theme_of_event(A, U),
  neothers(U, Nobel Peace Prize)

- **Semantic Constraints**. The environment tries to represent the whole semantic content of the question. Not all the items belonging to the environment are useful, however, for extracting the answer. So, depending on the QT, a subset of the environment has to be extracted. Sometimes additional relations, not present in the environment, are used and sometimes the relations extracted from the environment are extended, refined or modified. We define in this way the set of relations (the semantic constraints) that are supposed to be found in the answer. These relations are classified as mandatory or optional. Following the preceding example:

    - Mandatory Constraints:
        - actor_of_action(A, X)
        - action(A, won)
        - theme_of_event(A, U)
        - neothers(U, Nobel Peace Prize)
    - Optional Constraints:
        - time_of_event(A, T)
        - year(T, 1991)

- **Question Keywords.** The terms extracted from the question text that have to be used for performing the PR task.

Passage retrieval

In order to perform the PR task we have used MG [Witten et al, 1999]. Before the TREC-12 competition we indexed into MG the whole Acquaint collection. We built two indexes:

- Textual, i.e. indexing the documents from their textual content, as simple bag of words, with no pre-process
- Named Entities: We carried out a NERC process of the whole collection, we clustered these NE into clusters trying to group the different variants of the same entity, including acronyms for NE of type organization (see [Ferres et al, 2003b] for details of this process), and we indexed the documents using as key words the terms representative of the clusters.

At PR phase the process was the following:

With the Question Keywords obtained in the previous subsystem for each question, we looked for relevant documents in the two collections. We use a ranked retrieval for the textual collection with a threshold of 10% of the first retrieved document and a limit of 200 documents. For the Named Entities collection, we use Boolean retrieval in order to covers all the Named Entities of the query terms. The union of both sets of documents was indexed again into MG, this time only with the textual form but at level of passage. We consider a passage a sequence of 8 consecutive sentences of the original document allowing an overlapping of one sentence. Then, we retrieved the candidate passages with the same

keywords, again with a threshold of 10% of the first retrieved passage but this time using a limit of 50 passages.

Answer extraction

The process of extraction of the answer is carried out on the set of passages obtained from the previous subsystem. These passages are segmented into sentences. Each sentence is then scored according to its semantic content regarding the question. We build a general semantic representation of the concepts that occurring in the question in order to overcome the term-based approach limits for the sentence selection. The semantic content of a term is represented using a weighted vector, the weight of each term is computed using the *idf* of the term, synonyms, hyponyms and hyperonyms. The semantic content of a concept is then built from the semantic content of its terms. [Vicedo, 2002].

The linguistic process of extraction, quite expensive, is carried out on the sentences best scored.

This process is similar to the process carried out on questions and leads to the construction of the environment of each candidate sentence. The rest is a mapping between the semantic relations contained in this environment and the Semantic Constraints extracted from the question. The mandatory restrictions must be satisfied to take in consideration the sentence, the satisfaction of the optional constraints simply increases the score of the candidate.

The final extraction process is carried out on the sentences satisfying this filter.

The Knowledge Source used for this process is a set of extraction rules owning a credibility score. Each QT has its own subset of extraction rules that leads to the selection of the answer.

The process of application of the rules follows an iterative approach. In the first iteration all the semantic constraints have to be satisfied by at least one of the candidate sentences. If no sentence has satisfied the constraints, the set of semantic constraint is relaxed by means of structural or semantic relaxation rules, using the semantic ontology. If no candidate sentence occurs when all possible relaxations have been performed the question is assumed to have no answer.

Each candidate to solution comes weighted by diverse factors (sentence score, confidence of the used rules, satisfied optional restrictions, etc.). In the case more than one candidate is detected, a final process of weighted voting is carried out to select the preferred answer.

3. Evaluation and Conclusions

As we have said in the introduction, this paper describes the system we have built for our first participation in TREC competitions. Our main goal on attempting to participate in TREC-12 has been to acquire some experience on the kind of problems that have to faced in Q&A tasks. Although some of these problems have been foreseen by analysing other systems and previous competitions it is necessary to face the real problems (in real time) for taking the appropriate conclusions.

We have participate in TREC-12 with a prototype, not with a complete Q&A system. The different components of the system have got different levels of development (and, obviously, different level of accuracy). The first two subsystems, QP and PR, are the most completed and present the best results but the last one, AE, was only sketched and is currently under construction. Only a few rules for each Question Type were developed and no sufficient experimentation of the performance of these rules was carried out at the time of the competition.

With these constraints, the results obtained by our system are not good, but we think that they are a good starting point for further improvements of our system.

Some initial analysis of the results has been made and some comments follow.

As has been pointed above our participation was constrained to the factoid questions. So we provided answer to 413 questions. From them we gave the exact answer to only 22 questions (2 other were considered wrong). So the global accuracy of our system was 5.3%. The precision of recognising questions with no answer was of 9.2%, the recall was in this case of 43.3% (this figure was due to the fact that when our system was unable no find an answer the response was NIL, this was the case of 141 questions).

The classification of the question was rather good. The accuracy of our system was in this issue of 69%, quite close to the scores measured in our tests on previous TREC. Taking into account the fine granularity of our Types of Questions we think that it is a good result.

There 383 questions for which an answer exists in the collection. From these, only 157 were in the passages (50 per question) retrieved from the PR subsystem. This means that only 38% of the questions could be correctly answered by the AE subsystem.

As has been noted above, the AE module was only sketched for the competition. The accuracy of these components for questions which the answer occurred in our selected passages was of 8.3%.

Obviously, considered in isolation, the figures of AE are the worst of the three components. Part of the reason could be the accumulation of errors from previous components but the component itself has to be improved heavily.

There is of course enough room for improvement in every component and work is currently being done in all these components. In the next future, however, we plan to concentrate on getting better extraction rules by means of applying Machine Learning techniques, in the same line we have applied to the classification of questions problem.

References

[Atseries et al, 1998] J. Atseries, J. Carmona, I. Castellón, S. Cervell, M. Civil, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taules, J. Turmo
Morphosyntactic Análisis and Parking of Unrestricted Spanish Text.
Proceedings of First International Conference on Language Resources and Evaluation. LREC-98, Granada, Spain.

[Brants, 2000] Brant, T.
TNT, A Statistical Part-of-Speech Tagger,
Proceedings of the 6th ANLP-NAACL, Seattle, USA. 2000

[Carreras et al, 2002] Carreras, X., Màrquez, L. and Padró L.
Named Entity Extraction Using Adaboost.
Proceedings of the 6th conference on Computational Natural Language Learning (CoNLL 2002). Shared Task Contribution. Taipei, Taiwan. September 2002.

[Ferres et al, 2003a] D. Ferrés, M. Massot, M. Padro, H. Rodríguez, J. Turmo
Automatic Classification of Geographic Named Entities
4th LREC, 2004, Lisboa, Portugal. Submittted

[Ferres et al, 2003b] D. Ferrés, M. Massot, M. Padro, H. Rodríguez, J. Turmo
Automatic Building Gazetteers of Co-referring Named Entities
4th LREC, 2004. Lisboa, Portugal. Submittted

[Lin, 1998] Lin, D.
Dependency-based evaluation of MINIPAR
Proceedings of the Workshop on the Evaluation of Parsing Systems, LREC 1998, Granada, Spain

[Magnini, Cavaglià, 2000] B. Magnini, G. Cavaglià
Integrating Subject Field Codes into WordNet.
Proceedings LREC, Athens, Greece, 2000.

[Quinlan, 1993] Quinlan, J. R.
FOIL: A midterm report. In Proc. of the sixth European Conf. on Machine
Learning, Springer-Verlag, 1993.

[Rodriguez et al,1998] H. Rodriguez, S. Climent, P. Vassen, L. Bloksma, W. Peters, A. Alonge, F. Bertanga, A. Roventini
The Top-Down Strategy for Bulding EuroWordNet: Vocabulary Coverage, Base Conceps and Top Ontology.
Computer and Humanities 32. 1998, Kluwer Academic Publishers.

[Vicedo, 2002] J.L. Vicedo
Un modelo semántico aplicado a los sistemas de B´squeda de Respuestas.
Ph.D. thesis, Dept. LSI. Universidad de Alicante, May 2002.

[Witten et al, 1999] I. Witten, A. Moffat, T. Bell
Managing Gygabytes.
1999, Morgan Kauffman, San Francisco, second edition.