The PISAB Question Answering System

Giuseppe Attardi and Cristian Burrini
Dipartimento di Informatica
Università di Pisa - Italy
{attardi,burrini}@di.unipi.it

Abstract

The PISAB Question Answering system is based on a combination of Information Extraction and Information Retrieval techniques. Knowledge extracted from documents is modeled as a set of entities extracted from text and by relations between them.

During the learning phase we index documents using the entities they contain. In the answering phase we exploit the index previously built in order to focus the search for the answer to just the most relevant documents. As answers to a question we select from these documents the paragraphs containing entities most similar to those in the question.

PISAB has been submitted to the TREC-9 Conference, achieving encouraging results despite it current prototypical development stage.

Introduction

The problem of finding answers to questions on a large document collection, could in principle be solved by creating a knowledge base with the information extracted from documents and then querying such knowledge base. Unfortunately this approach is not yet feasible, since it requires advanced techniques of natural language processing, knowledge extraction, knowledge representation and reasoning, which are beyond the current state of the art.

On the other hand, Information Retrieval techniques are quite effective in retrieving documents relevant to a certain subject, so in particular those which might contain the answer to a question. Information Extraction techniques help identifying certain kinds of information, but their capabilities are quite domain dependent and limited to entities with predefined patterns. Neither of these techniques is sufficient to address the Question Answering problem, but we have explored a way of combining them to build a complete question answering system.

The approach

The meaning of a document might be expressed in terms of entities and relations between them. Entities are the semantic equivalent of nouns present in the document, while relations correspond to verbs. For example, the information contained in the following phrase:

"John reads a book"

can be represented by means of the entities "John" and "book", and by the relation "reads" that links subject and object. Relations need not be binary: prepositional phrases and various kinds of syntactic adjuncts allow expressing *n*-ary relations.

Intuitively an entity refers to an object in the real world or to a concept, or in general to an element of the semantic domain of the document contents. In this semantic domain entities have attributes and are related to other entities [Attardi 86]. Within documents only syntactic representations of the entities are present, expressed in one of several forms:

- Proper nouns: **Microsoft**;
- Pronouns reference: Microsoft … **It**;
- Descriptions in terms of attributes values: **the world largest software company**;
- Descriptions in terms of relations among entities: **the Redmond company**.

Suitable natural language processing techniques are available for extracting entities expressed in the first two ways. The case of entities referenced by proper nouns is referred in the literature as *Named Entity extraction*. The case of pronouns is called *coreference* and *anaphora resolution*. Techniques for dealing with objects expressed by means of attributes or relations descriptions are less common.

Our system tries to locate the syntactic boundary of each kind of entity expression and, in the first two cases, it attempts to solve the references. The architecture of our system consists of an *Entity Tagger*, which performs a superset of the functions of a Named Entity tagger, and a *coreference module* (of still limited capabilities in our prototype).

Entity Tagger

The Entity Tagger works mainly at the syntactic level, exploiting features provided by lexical analyzers, like part of speech tags, cases, gazetteers and contexts rules.

For example, given the phrase “The TREC-9 Conference was held in Maryland in November 2000” the Entity Tagger produces:

< TREC-9 Conference> was held in <Maryland> in <November 2000>.

The Entity Tagger is capable of dealing with **structured entities**. Numeric dates, Web addresses, numbers, monetary expressions are examples of structured entities, which are expressed according to specific syntactic rules. Suitable Information Extraction modules are invoked in a pre-processing stage of the Entity Tagger in order to recognize structured entities.

Semantic Tagger

Having determined the syntactic boundaries of an entity, the system tries to classify it according to a predefined semantic ontology, i.e. it tries to assign to each expression the proper semantic category of the referred object.

For example, given the phrases:

<Washington> is in <North America>.

<George Washington> didn’t like <apples>.

<Washington> threatens <Iraq> to start <the war>.

the semantic tagger is capable of disambiguating the three senses of the term “Washington” within each phrase, classifies each occurrence accordingly, and produces:

[Washington/LOCATION] is in [North America/LOCATION].

[George Washington/PERSON] didn't like [apples/FOOD].

[Washington/ORGANIZATION] threaten [Iraq/ORGANIZATION] to start [the war/ACT].

To perform classification, the semantic tagger exploits a thesaurus and a set of *context rules*, which allow inferring the semantic category of any entity (not just of a Named Entity). The semantic ontology used for classification is a two level tree extracted from the *is-a* relation of the WordNet [WordNet] thesaurus.

$$\begin{gathered} Chunk_1{:}T_1,\ \ Chunk_2{:}T_2,\ \dots\ Chunk_n{:}T_n \\ AND \\ Attribute\text{-}Conditions(Chunk_1,..,Chunk_n) \\ \rightarrow add\text{-}Score(Chunk_k,sem\text{-}Vector) \end{gathered}$$

Figure 1 - Structure of a Context Rule

$$\begin{gathered} C_1{:}NamedEntity,\ C_2{:}Verb,\ C_3{:}Entity \\ AND \\ (C_2.HeadWord.rootForm=\text{"be"}\ AND \\ C_3.semantic\_approximation=semx) \\ \rightarrow add\text{-}Score(Chunk_1,semx) \end{gathered}$$

Figure 2- The is-a context rules

In its general form, a context rules states that if the context surrounding an entity matches a certain syntactic pattern and certain conditions on semantic attributes of the entities involved are met, then the context provides evidence that the entity belongs to a certain semantic category, expressed as a numeric weight. For example, the rule in Figure 2, applied to fragment "John is the chief director of …", states that the entities <John> and <the chief director> most likely have similar meanings. Entity classification is probabilistic based: the semantic tagger computes the likelihood that an entity belongs to each semantic category, and selects the category with the highest likelihood.

The pair of an entity instance and its semantic category makes what we call a *concept*, for instance ⟨George Washington, PERSON⟩.

Concept Indexing and retrieval

Our QA system is based on a concept-oriented indexing and search engine which stores the concepts extracted from documents using IE techniques. The underlying intuition is that both the documents relevant to a question and the question itself must be about the same semantic objects. Therefore documents are indexed according to the semantic entities extracted from them, and they are searched then through to the semantic entities extracted from queries. We expect the search will match only a small set of documents, strictly related to the question, within which to focus further search for the answer, since it is most likely that they contain such answer.

This document-filtering step has also efficiency benefits, since it reduces the number of documents that must be considered in subsequent steps.

Architecture

The overall architecture of PISAB Question Answering System is illustrated in Figure 3:

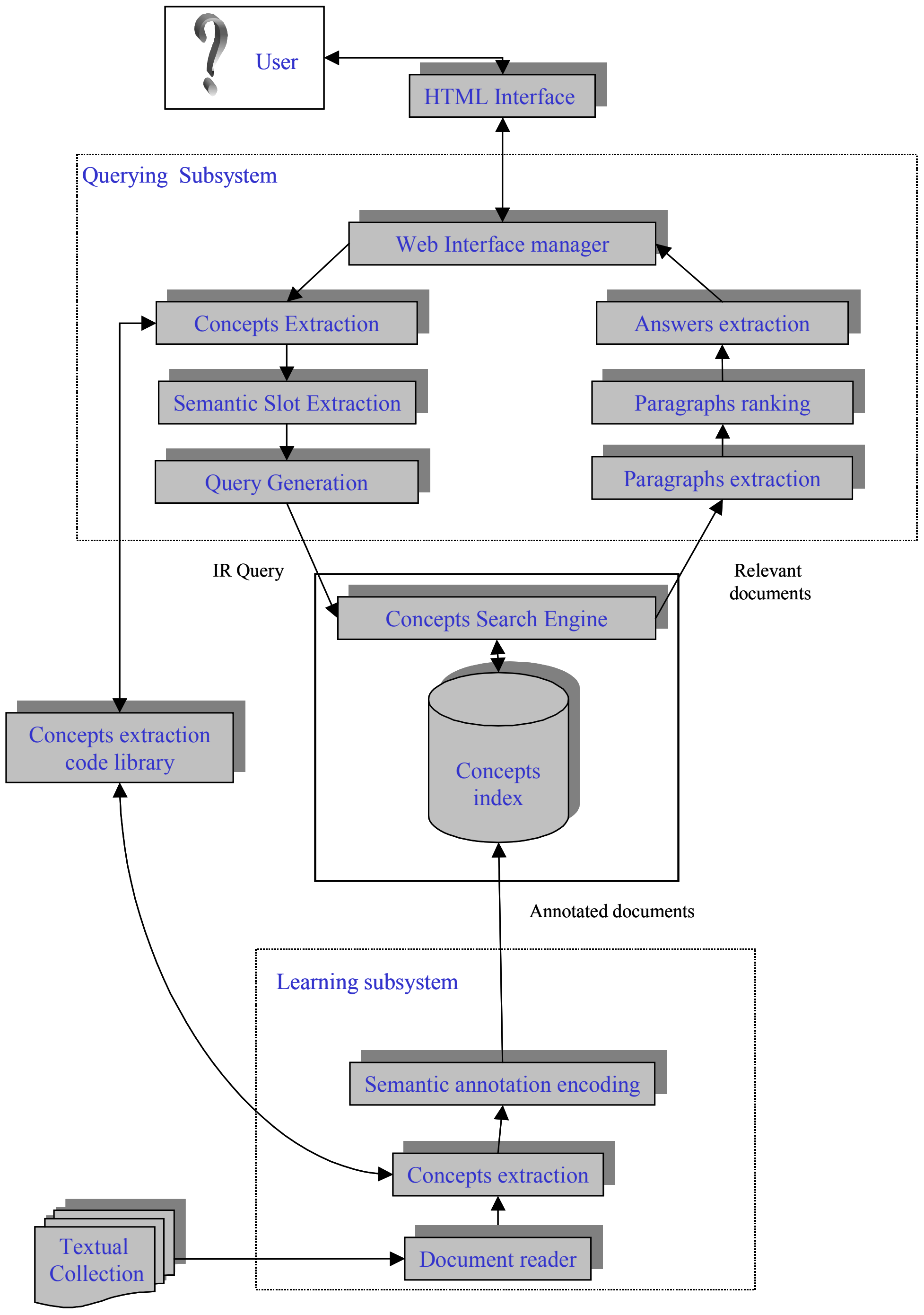


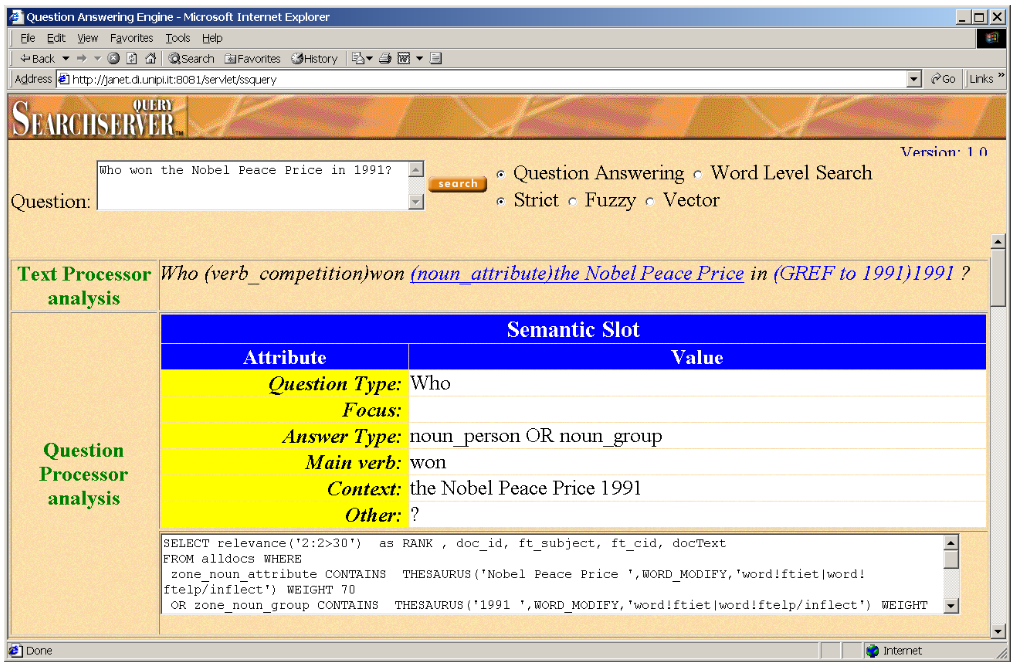
Figure 3 – Architecture of PISAB Question Answering System.

The learning subsystem extracts concepts from documents (i.e. semantic classified entities), by means of the semantic tagger. Concepts are used as indexes for the analyzed documents in a document retrieval system.

The first step in the answering phase is to extract concepts from the question. Then a query for the document retrieval system is constructed, made up from these concepts, but with different weights assigned to each concept according to its role in the question. For example, a concept that seems to be the focus of the question is weighted more than one that is not. (The focus is a concept that describes semantically the answer entities. For example: in “What U.S. President did …” the *Question Focus* is “U.S. President” and the *AnswerType* is PERSON).

This is done by analyzing the question and building an abstract representation of it consisting of several *semantic slots*, including:

- Question type: who/where/when/which/how, …
- Main verb
- Description of the concept to find (*Focus*)
- Semantic category of the candidate answer (*Answer type*)
- Other contextual concepts (*Context*)


Question Answering Engine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://janet.di.unipi.it:8081/servlet/ssquery
Go Links
QUERY SEARCHSERVER
Version: 1.0
search
Done
Internet

Question: Who won the Nobel Peace Price in 1991?

[✓ Question Answering] Word Level Search

[✓ Strict] Fuzzy Vector

Text Processor analysis: *Who (verb_competition)won (noun_attribute)the Nobel Peace Price in (GREF to 1991)1991 ?*

Question Processor analysis

Semantic Slot	
Attribute	**Value**
Question Type:	Who
Focus:	
Answer Type:	noun_person OR noun_group
Main verb:	won
Context:	the Nobel Peace Price 1991
Other:	?

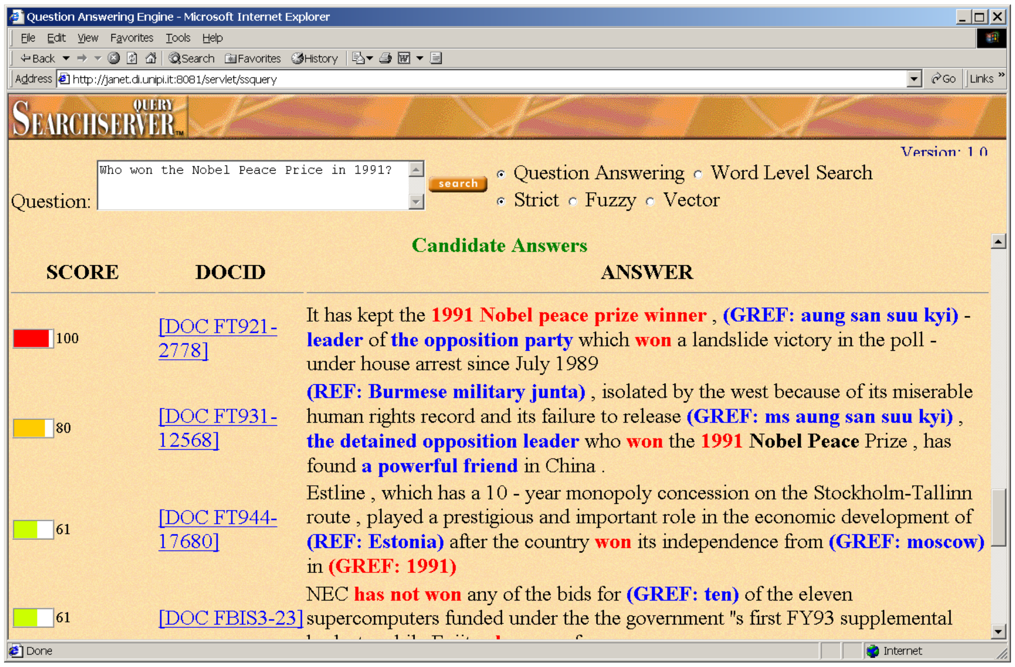
```
SELECT relevance('2:2>30')  as RANK , doc_id, ft_subject, ft_cid, docText
FROM alldocs WHERE
 zone_noun_attribute CONTAINS  THESAURUS('Nobel Peace Price ',WORD_MODIFY,'word!ftiet|word!
ftelp/inflect') WEIGHT 70
 OR zone_noun_group CONTAINS  THESAURUS('1991 ',WORD_MODIFY,'word!ftiet|word!ftelp/inflect') WEIGHT
```

Figure 4 PISAB prototype Web Interface.
Shown are: question, choice of IR search method, results of NLP analysis of question, filled semantic slots and sketch of generated query.

To increase the robustness of document relevance ranking we combine concept weighting with a traditional IR scoring function: the *cosine distance* between the text and the query. Adding such term-related score mitigates the effects of errors in entity extraction and classification. The query is finally expanded by means of a thesaurus, including variants of concepts and terms present in the question, for improving matching likelihood.

The expanded query is used to retrieve documents relevant for the question. The most relevant ones, according to the search engine rank, are further analyzed in order to extract the candidate answers. We split the document into paragraphs (or sentences) and rank them according to an estimate of relevance to the query. We have a concept occurrence, also called a “semantic hit”, when a paragraph contains an inflection of a question’s concept or an entity with the same semantic category of the Question Focus Each occurrences is weighted according to the associated semantic slot and all weights are added up to obtain a score for each paragraph.

Finally, from each top ranked paragraph we extract the best scoring text window: this will be one of the candidate answer to be displayed to the user, as shown in Figure 5.


Question Answering Engine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://janet.di.unipi.it:8081/servlet/ssquery
Go Links
QUERY SEARCHSERVER
Version: 1.0
search
Done
Internet

Question: Who won the Nobel Peace Price in 1991?

[✓ Question Answering] Word Level Search
[✓ Strict] Fuzzy Vector

Candidate Answers

SCORE	DOCID	ANSWER
🟥 100	[DOC FT921-2778]	It has kept the **1991 Nobel peace prize winner** , **(GREF: aung san suu kyi)** - **leader** of **the opposition party** which **won** a landslide victory in the poll - under house arrest since July 1989
🟧 80	[DOC FT931-12568]	**(REF: Burmese military junta)** , isolated by the west because of its miserable human rights record and its failure to release **(GREF: ms aung san suu kyi)** , **the detained opposition leader** who **won** the **1991 Nobel Peace** Prize , has found **a powerful friend** in China .
🟨 61	[DOC FT944-17680]	Estline , which has a 10 - year monopoly concession on the Stockholm-Tallinn route , played a prestigious and important role in the economic development of **(REF: Estonia)** after the country **won** its independence from **(GREF: moscow)** in **(GREF: 1991)**
🟨 61	[DOC FBIS3-23]	NEC **has not won** any of the bids for **(GREF: ten)** of the eleven supercomputers funded under the the government "s first FY93 supplemental

Figure 5 – Candidate Answers.

References

[Attardi 86] G. Attardi and M. Simi, A Description Oriented Logic for Building Knowledge Bases, *Proc. of the IEEE*, Vol. 74, N. 10, 1986, 1335-1344.

[Lasso] D. Moldovan, S. Harabagiu, M. Pasca, R.Mihalcea, R. Goodrum, R. Girju, V. Rus. Lasso: A Tool for Surfing the Answer Net. Department of Computer Science and Engineering Southern Methodist University. Dallas, 1999.

[LTG] A. Mikheev, C. Grover, M. Moens. Description of the LTG system used for MUC-7. HCRC Language Technology Group, University of Edinburgh, 1998.

[PACLING] S. Baluja, V.O. Mittal, R.Sukthankar. Applying machine learning for high performance Named-Entity extraction. PACLING'99, Waterloo, Canada, 1999.

[Textract] Rohini Srihari, Wei Li. Question Answering Supported by Information Extraction. Cymfony Inc, Williamsville CA, 1999.

[TREC8] Ellen Voorhess, Dawn M. Tice. The TREC-8 Question Answering Track Evaluation. National Institute of Standards and Technology, Gaithersbugh, 1999.

[VIE] K. Humphreys, R.Gaizauskas, H. Cunningham, S. Azzam. VIE Techincal Specification. Department of Computer Science and Institute for Language, Speech and Hearing (ILASH) University of Sheffield, 1998.

[WordNet] G.A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41, 1995.