Mining Textual Answers with Knowledge-Based Indicators

Sanda M. Harabagiu and **Marius Paşca**
Department of Computer Science and Engineering
Southern Methodist University
Dallas, TX 75275-0122
{sanda,mars}@seas.smu.edu

Abstract

This paper describes a knowledge-based methodology of mining textual answers into large collections of texts. We present **SOMBRERO**, a knowledge processing module implemented in the **LASSO** question answering system. The module bootstraps question taxonomies, used for the search and validation of answers. Answer evaluation is produced through an expandable set of abduction rules.

Background

When people use computer-based tools to find answers to general questions, they often are faced with a daunting list of search results, or "hits" returned by a search engine. To take a step closer to *information retrieval* rather than *document retrieval*, this year the Text REtrieval Conference (TREC) has initiated an experimental track: the *Question/Answering* (Q/A) track, whose aim is to foster research in the area of textual Q/A, especially when pursued in a domain independent manner. The TREC Q/A Track specifies two restrictions for the questions. First, questions should have an exact answer that occurs in some document from the underlying text collections[1]. The second restriction applies to the length of the answer. There are two answer categories. The first one limits the length to 50 contiguous bytes of text whereas the second category comprises answers that either (a) represent a sentence or (b) are under 250 bytes of text.

As participants in the TREC-8 Q/A competition, we contributed to the design of **LASSO**, a Q/A system that combines *paragraph indexing* with *lightweight abduction* of answers, cf. (Moldovan et al.1999). **LASSO** performed surprisingly well in the TREC competition, achieving 55.5% precision for the short answer category and 64.6% for the long answers. However, we have noticed that even when the correct answer was not the first one returned, it still could be found in one of the top five answers in 68.1% of the short answers and in 77.7% of the long answers. This shows that our empirical abductive methods are not strong enough to evaluate the correctness of an answer. They are based on the information an answer *wears on its sleeve*, available via surface-text-based methods. Thus, we realized that higher precision cannot be achieved unless we employ richer knowledge structures, enabling stronger mechanisms for weighted abduction.

With this goal in mind, we set to develop **SOMBRERO**, a knowledge-processing module that ports semantic knowledge and the results of abductive reasoning to the **LASSO** Q/A system. Knowledge processing at both question and answer level is performed. Moreover, both specific knowledge bases and abduction rules are acquired in a meta-boosting context. In the remaining of the paper, we describe the architectures of **LASSO** and **SOMBRERO** and we detail our question ontology as well as the abductive mechanisms. Results of our evaluations are also presented, as well as plans for future work.

Knowledge Processing for Q/A

To find the answer of a question by returning a small fragment of text, where the answer actually lies, a Q/A system needs to combine an indexing scheme with knowledge processing of the question and of the answer.

Initially, the *Question Processing* module developed in **LASSO** transformed a question of open-ended nature, expressed in natural language into a three-featured representation of the information requested. This representation comprised (a) a *question type*, selected from a manually generated classification, (b) the *answer type*, indicated by semantic dictionaries and (c) the *question focus*, defined as the main informa-

[1] In TREC-8, the text collections have comprised 131,896 *LA Times* articles, 130,471 *Foreign Broadcast Information Service* messages, 55,630 articles from the *Federal Register* and 210,158 articles from the *Financial Times*.

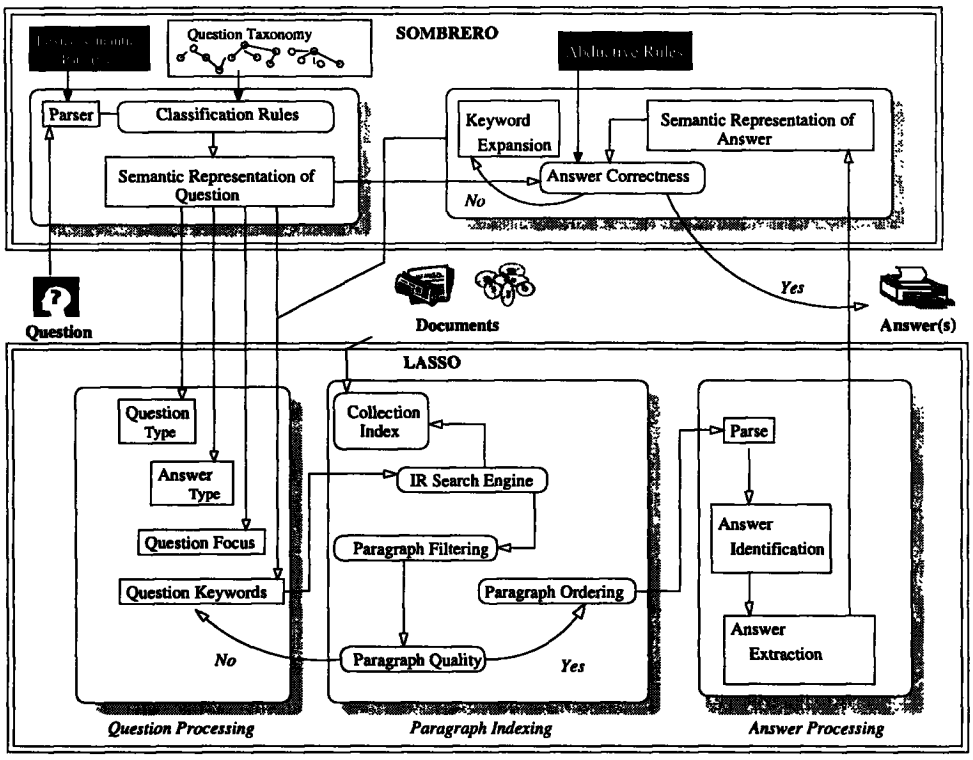


Figure 1: SOMBRERO: A knowledge processing module for the LASSO Q/A System

tion required by the interrogation. Furthermore, the *Question Processing* module from **LASSO** identifies the *question keywords* that serve in the retrieval of the answers. All the *Question Processing* performed in **LASSO** is of empirical nature.

To be able to address a larger variety of question classes and especially to provide a richer semantic knowledge representation, we have developed **SOMBRERO**, a knowledge processing module that generates the question type, the answer type, the question focus and the question keywords based on a semantic representation of the question. As illustrated in Figure 1, first the question is parsed and lexico-semantic patterns are applied to detect semantic regularities. Then, the question is classified against a taxonomy of questions, generating the semantic representation of the question. This representation will also be used in the evaluation of the correctness of the answer.

Based on the question keywords, the *Paragraph Indexing* module from **LASSO** produces a set of ordered paragraphs, in which all the keywords were found by boolean matching. Paragraph indexing is one of the major innovations of **LASSO**, since it returns paragraphs of interest instead of documents of interest. Moreover, when the paragraph quality is not acceptable, another set of keywords is submitted to the search engine, producing an Information Retrieval (IR) loop that increases the overall quality of the retrieval. However, from our experiments, we have determined that the IR precision and the overall performance of the Q/A system is influenced in large measure by the quality of the keywords. To address this issue, **SOMBRERO** has generates keywords by using the semantic representation of the question, thus improving the precision of **LASSO**.

The *Answer Processing* module parses the paragraphs that contain the question keywords, and identifies answer candidates, based on the answer type. Answers are then extracted, by using shallow abductive mechanisms, working on bag-of-words approaches. The results of **LASSO** in the TREC-8 Q/A competition and some of our experiments have indicated that greater precision can be achieved if answer correctness could be evaluated.

In **SOMBRERO**, we translate the candidate answer into the same semantic representation as the question. A set of abductive rules controls the unification of the two representations. We implement a weighted abduction, that generates scores of answer correctness. When the abduction is successful, the answer is produced. Otherwise, keyword expansion of unmatched terms takes place, and the retrieval resumes. If this

second retrieval loop does not improve the correctness scores, the search for an answer is abandoned.

A Question Taxonomy

Our question taxonomy is based on two principles. First, we used a top-level categorization and a set of constraints that define a seed hierarchy of questions. On top of this, we define a new representation of each question, defined by a pair of a (a) generally ambiguous question stems (e.g. what, where) (b) a semantic representation. Second, based on this representation, we learn classification rules, as well as new classes of questions.

The top-level categorization is provided by the taxonomy introduced in (Graesser et al.1992). This classification has the advantage that it has been empirically evaluated for completeness. The seed taxonomy was further refined by constrains defined in QUALM, the computational theory of Q/A reported in (Lehnert 1978). The major question classes are: (1) *Verification* (e.g. Is a fact true? Did an event occur?); (2) *Comparison* (e.g. How similar/ different is X from Y?); (3) *Disjunctive* (e.g. Is X or Y true?); (4) *Concept completion* (e.g. Who...? What...? When...? Where...? Which...?); (5) *Definition/Example* (e.g. What are the properties of X?); (6) *Quantification* (e.g. How much...? How many...?); (7) *Feature specification* (e.g. What is the attribute of X?); (8) *Causal antecedent/consequent* (e.g. What caused some event?); (9) *Procedural* (e.g. What instrument enables an act?) (10) *Expectational/Analogical* (e.g. Why did some event not occur?) and (11) *Judgemental* (e.g. Request for a judgemental/advise/interpretation).

Our contribution to the taxonomy is determined by (a) a new representation of questions and (b) a meta-bootstrapping mechanism that allows to learn both new question classes and new classification rules at the same time. Our representation combines *lexical elements* with a semantic structure that relates the *answer type* imposed by the meaning of the question with the *question focus* and the other concepts of the question. Figure 2 illustrates some of the representation used in the *Concept completion* class of questions, which represents the most numerous sub-hierarchy.

In Figure 2, we illustrate three subclasses, the *NAME* questions, that look for an answer represented as a named entity, the *PRODUCT* class, associated with answers that represent artifacts or other creations and the *MONEY* class, returning answers that a numeric values of different kinds of currencies. For a question type, we may have multiple answer types. For example, when asking about a named entity, we may ask about a person, a product, an organization,

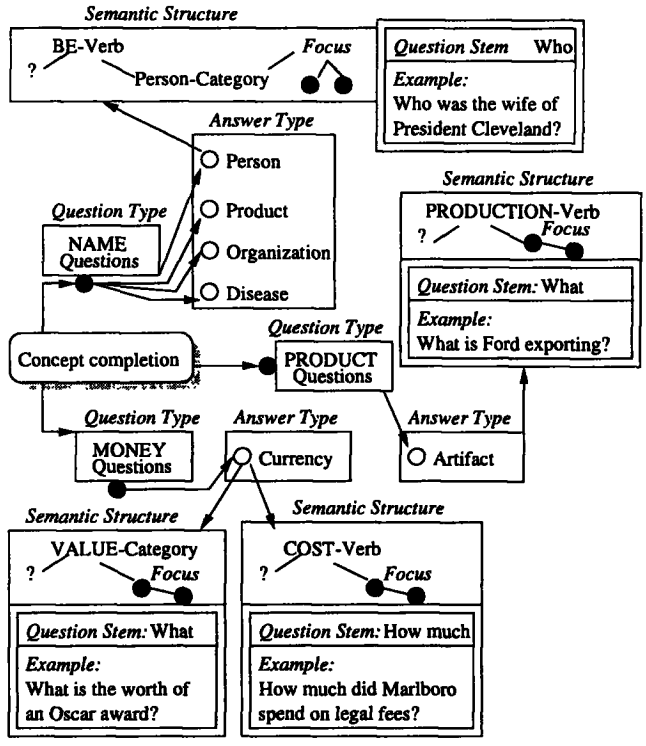


Figure 2: A snapshot from the Question Taxonomy

a disease or a location. In turn, to each answer type, we associate multiple semantic representations of questions. A semantic representation in **SOMBRERO** is a case frame with anonymous relations, that would allow the unification of the answer and of the question regardless of the case relation. Two case slots have special meaning in this representation: (1) the answer type, represented with a question mark in the question frame, and unifiable with the answer type recognized in a text, and (2) the focus, whose position is defined as an adjunct or parallel of the answer type in the syntactic parse. The keywords are all syntactic adjuncts of the *question focus.*

A set of classification rules is associated with this taxonomy. Initially, all rules are based on the recognition of the *question stem* and of the *answer type*, obtained with class information from WordNet. However we could learn new rules when we allowed for morphological and semantic extensions of the terms from the semantic representations of the initial taxonomy. Along with the new rules, we enriched the taxonomy, as a result of new questions being unified only partially with the current taxonomy. Moreover, the training data is easily accessible, since there are a large number of FAQ (frequently asked questions) available on the Internet.

The bootstrapping algorithm that allows to learn new classification rules and new classes of questions is based on an information extraction measure: $score(rule_i)=A_i * log_2(N_i)$, where A_i stands for the number of different lexicon entries for the *answer type of the question*, whereas $N_i = A_i/F_i$, where F_i is the number of different focus categories classified. The steps of the boostrapping algorithm are:

1. Retrieve concepts morphologically/semantically related to the semantic representations
2. Apply the classification rules to all questions that contain any newly retrieved concepts.
3. New_Classificatiton_Rules={}
MUTUAL BOOTSTRAPPING LOOP
4. Score all new classification rules
5. best_CR=the highest scoring classification rule
6. Add best_CR to the classification rules
7. Add the questions classified by best_CR to the taxonomy
8. Goto step 4 three times.
9. Discard all new rules but the best scoring three.
10. Goto 4. until the Q/A performance improves.

Answer Processing

In (Hobbs et al.1993) a method of abductive interpretation of texts was introduced, describing its operation in the TACITUS system. This method imposes two challenging conditions: (a) exhaustive world knowledge representation and (b) the capability of limited, efficient backward chaining. For a Q/A system, these requirements are impractical. However, the need for answer abduction is determined by the limited performance of bag-of-words approaches to answer extraction.

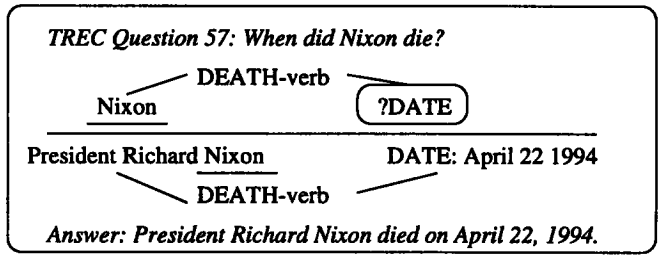
In **SOMBRERO** we provide with several abductive rules that mine successfully the correctness of answers from the knowledge available at hand. Two factors have impacted the precision enhancement of **LASSO**. First, the question taxonomy provides with correct recognition of the question focus. In **SOMBRERO** we discard all candidate answers that do not contain the question focus. Only this requirement has impacted on the precision of **LASSO** in two major ways: (a) correct answers that were not top-ranked in **LASSO** now become the first answer returned in 18 of the cases; (b) 7 questions that did not have correct answers were solved successfully. Table 1 illustrates two such examples of TREC-8 question whose answer is improved **SOMBRERO**. The answer of question 26 was initially placed on the second position (out of five), being incorrectly preceded by an answer that did not contain the focus. In the case of TREC question 198, no correct answer was initially found. With the means of focus detection, the correct answer is retrieved. In both examples, the question focus is underlined.

Question-26	What is the name of the "female" counterpart of El Nino, which results in cooling temperatures and very dry weather?
Answer (short)	Score: 416.00 *The Times believes that the greenhouse effect and El Nino- as well as its "female" counterpart, La Nina - have had a profound effect on weather*
Question-198	How did Socrates die?
Answer (long)	Score: 144.00 *refute the principle of retaliation that Socrates , who was sentenced to death for impiety and the corruption of the city ' s youth , chose to drink the poisonous hemlock, the state ' s method of inflicting death*

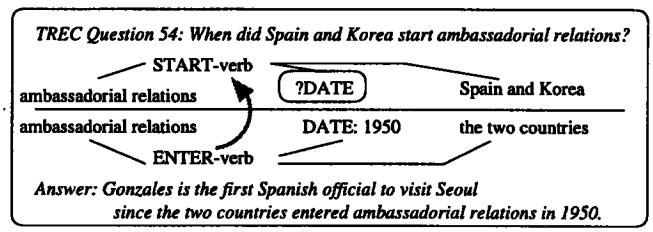
Table 1: Examples of improved answer correctness.

The second impact of the question taxonomy on the evaluation of answer correctness comes from the unification of the semantic representations of the question and the answer. Our abduction rules enable this unification, by providing approximations of the semantic information. In fact, we perform a weighted abduction, that takes into account semantic distance between question and answer concepts. Some of the abductive rules are:

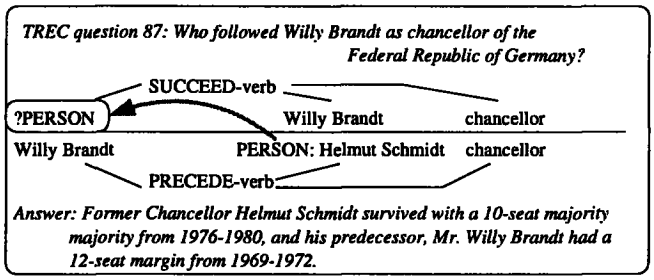
• *Abduction Rule 1.* If the unification of the question and answer representations comprises the focus, if one of the concepts in the answer disambiguates the corresponding question concept, assign a score equal to # identical concepts/ # unifiable concepts. An example is:



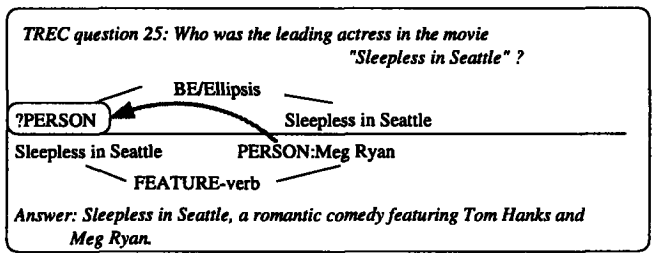
• *Abduction Rule 2.* If the unification of the question and answer representations involves temporal inclusion or overlap of the verb controlling the focus, assign a score equal to # identical concepts + # concepts controlled by the temporal event/ # unifiable concepts. An example is:



• *Abduction Rule 3.* If the unification of the question and answer representations involves an asymmetrical relation assign a score equal to # identical concepts + # concepts controlled by the relation/ # unifiable concepts. An example is:



• *Abduction Rule 4.* If the unification of the question and answer representations involves several relations and concepts in the answer that account for the ellipsis phenomenon in the question (i.e. they are not mentioned, but implied) then assign a score equal to # identical concepts / # unifiable concepts + # relations controlled by ellipsis. An example is:



The bootstrapping of abduction rules uses a similar algorithm as the one presented for the question taxonomy. It is characterized by combinations of the seed abduction rules. Currently we use a set of 18 seed rules.

Evaluation

We have evaluated the contribution of **SOMBRERO** on the performance of the **LASSO** Q/A system through five experiments. In the first two experiments we have used only the question taxonomy, without the support of the abduction rules. The other three experiments have used the abductive component. In the first experiment, we have employed only the seed question taxonomy, whereas in the second experiment we made use of the taxonomy generated by the bootstrapping process. In the third experiment, we have used no question taxonomy, but have employed the seed abduction rules, whereas in the fourth experiment we have employed the full set of abduction rules. In the final experiment, we have used both the bootstrapped taxonomy and the bootstrapped set of abduction rules. In these experiments, we had a taxonomy of 48 classes of questions and a set of 25 abduction rules. Figure 3 illustrates the impact on the precision of **LASSO** in each experiment. It is noticeable that the effect of finer question taxonomies has greater impact than larger number of abduction rules.

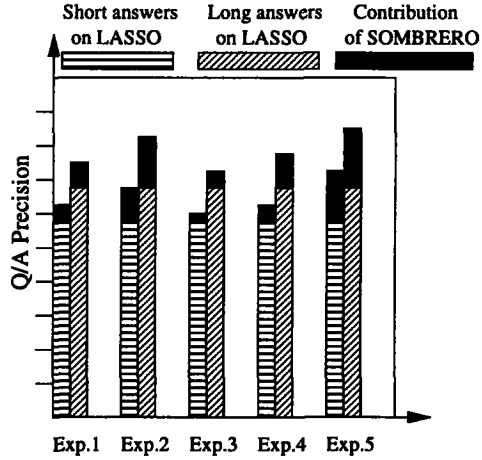


Figure 3: Experimental results on LASSO

Conclusions

We believe that the performance of a Q/A system depends on the knowledge sources it employs. In this paper we have presented the effect of knowledge derived from question taxonomies and abductive reasoning on Q/A precision. In the future, we plan to study the influence of additional knowledge on Q/A systems. Our study will evaluate the effect of knowledge derived from reference resolution and discourse coherence on Q/A precision.

References

Arthur Graesser, Natalie Person and John Huber Mechanisms that generate questions. In Lawrence Erlbaum Associates, *Questions and Information Systems*, pages 167–187., editors Lauer, T.W., Peacock,E. and Graesser, A.C., Hillsdale, N.J., 1992.

Sanda Harabagiu and Steven Maiorano. Finding answers in large collections of texts: paragraph indexing + abductive inference. *Working Notes of the Fall AAAI Symposium on Question Answering*, November 1999.

Jerry Hobbs, Mark Stickel, Doug Appelt, and Paul Martin. Interpretation as abduction. *Artificial Intelligence*, 63, pages 69–142, 1993.

Wendy Lehnert. The processing of question answering. Lawrence Erlbaum Publishers, 1978.

Dan Moldovan, Sanda Harabagiu, Marius Paşca, Rada Mihalcea, Richard Goodrum, Roxana Gîrju and Vasile Rus. Lasso: a tool for surfing the answer net. In *Proceedings of TREC-8*, 1999.