The "La Sapienza" Question Answering system at TREC-2006

Johan Bos
Dept. of Computer Science
University of Rome "La Sapienza"
bos@di.uniroma1.it

Abstract

This report describes the system developed at the University of Rome "La Sapienza" for the TREC-2006 question answering evaluation exercise. The backbone of this QA system is linguistically-principled: Combinatory Categorial Grammar is used to generate syntactic analyses of questions and potential answer snippets, and Discourse Representation Theory is employed as formalism to match the meanings of questions and answers. The key idea of the La Sapienza system is to use semantics to prune answer candidates, thereby exploiting lexical resources such as WordNet and NomLex to facilitate the selection of answers. The system performed reasonably well at TREC-2006: in the per-series evaluation it performed slightly above the median accuracy score of all participating systems.

1 Introduction

The QA evaluation exercise at TREC consists in automatically finding answers for a collection of questions arranged by different topics, or *targets* in TREC parlance. Questions can be either *factoid*-questions, asking for a unique short answer, or *list*-questions, asking for a set of answers. Each series of questions ends with an *other*-question, which is a request of providing all relevant information about the target which was not already asked in the previous questions. An example of a target and its questions is shown in Figure 1.

The answers must be found in the Aquaint corpus, a collection of over a million articles in English prose from three different newspapers, dating from 1996–2000. A response is evaluated as correct only if it exactly answers the question (in an exhaustive but not overinformative way) and if it is accompanied by a document ID from the Aquaint corpus supporting the answer.

This paper contains a description of the TREC-2006 entry of the University of Rome "La Sapienza" for the question-answering evaluation exercise. Probably the most interesting aspect of the La Sapienza system is that it is linguistically principled, combining symbolic with statistical approaches. The system is very similar to the QED system described in (Leidner et al., 2003; Ahn et al., 2004; Ahn et al., 2005), in that it also uses CCG (Combinatory Categorial Grammar) and DRT (Discourse Representation Theory) to generate meaning representations for questions and answer contexts, with the key idea that semantics helps to prune possible answer candidates.

2 The La Sapienza QA system

2.1 Question Analysis

The question is tokenised and parsed (together with the target) with the wide-coverage CCG-parser of Clark & Curran (Clark and Curran, 2004). On the basis of the output of the parser, a CCG-derivation, a semantic representation is constructed in the shape of a Discourse Representation Structure (DRS), closely following Discourse Representation Theory (Kamp and Reyle, 1993). This is done using the semantic construction method described in (Bos et al., 2004; Bos, 2005). The Question-DRS forms the basis for generating other pieces of information:

- an answer type, the answer cardinality, and tense (Section 2.2);
- background knowledge from lexical resources (see Section 2.3);
- a query for document retrieval (Section 2.4).

The idea is that the analysis of the question gives us all the information required later in the question answering process. For instance, not all of the available background knowledge is selected, but only those parts relevant for answering the question.

2.2 Answer Types

The La Sapienza system distinguishes fourteen main answer types (which are further divided into subtypes, but not discussed in this paper). The answer types play a role in extracting and selection of answers. The different types and examples of questions are shown in Table 1.

The answer cardinality denotes a range expressed by an ordered pair of two numbers, the first indicating the mininal number of answers expected (the lower bound),

TARGET: stone circles

169.1 (factoid) When did the construction of stone circles begin in the UK?
169.2 (factoid) Approximately how many stone circles have been found in the UK?
169.3 (factoid) When was Stonehenge built?
169.4 (factoid) In what county was Stonehenge built?
169.5 (list) What are the locations or names of other stone circles in the UK?
169.6 (factoid) What is the oldest stone circle in the UK?
169.7 (other)

Figure 1: Example of a TREC-2006 serie of questions for a target.

the second the maximal number of answers (the upper bound, which is set to 0 if unknown). For instance, 3–3 indicates that exactly three answers are expected, 2–0 means at least two answers. This information is used for determining the number of answers to be returned for list questions. Again, see Table 1 for examples.

Finally, the answer type is complemented with the tense in which the question is posed. This is a value of the set {past, present, future}. Currently this feature is only exploited for restricting potential answers that denote a temporal value. Once again, see Table 1 for examples.

2.3 Background Knowledge

The background knowledge for a question constitutes a list of axioms related to the question. It is gathered from lexical resources on the basis of the symbols that occur in the semantic representation of the question. Currently the following kinds of axioms are used:

- synonyms and hyponyms for nouns and verbs derived from WordNet (Fellbaum, 1998);
- hyponyms for nouns harvested from corpora using lexical patterns using techniques similar as in (Hearst, 1992);
- nominalisation rules generated from NomLex (Meyers et al., 1998);
- specialised knowledge, such as attributes (colours, shapes), and geographical knowledge (continents, states, countries, capitals);
- a couple of hand-crafted general inference rules.

The background knowledge for a question is used when extracting potential answers from contexts, and in the answer reranking.

2.4 Document Retrieval

All documents in the Aquaint corpus were pre-processed: the XML was stripped off, and the sentences were split and tokenised. The documents were rearranged into smaller documents of two sentences each (taking a sliding window, so each sentence appeared in two mini-documents). These mini-documents were indexed with the Indri information retrieval tools (Metzler and Croft, 2004).

Two kinds of queries are generated for each question: a complex query, based on the target and the information within the question; and a simple query which is just identical to the target. The different types of query were used in two different runs to find out whether one outperformed the other. It had been noticed already that simply using the target as query yields pretty good results (Ahn et al., 2005).

The best 1,500 mini-documents are retrieved, again with the help of Indri (Metzler and Croft, 2004). At this stage of processing, the aim is high recall at the expense of precision. By selecting a high number of documents, the pool of potential answers can be narrowed down as late as possible in the processing pipeline. Processing a high number of documents is certainly time-consuming, but since there are no important time-constraints in the TREC exercise, this is no big concern and advantage is taken of this situation.

2.5 Document Analysis

Using the same wide-coverage parser as for parsing the question, all retrieved documents are parsed and for each of them a Discourse Representation Structure (DRS) is generated. The parser also performs basic named entity recognition for locations, persons, and organisations (Curran and Clark, 2003). This information is used to assign the right semantic type to discourse referents in the DRS.

Each passage is translated into a single DRS; hence a DRS can span several sentences. A set of DRS nor-

Table 1: Answer type, cardinality, and tense for some examples of the TREC-2006 test set.

Answer Type	Card	Tense	ID	Example
DESCRIPTION	1–1	present	198.4	What is the claimed primary purpose of this facility?
ATTRIBUTE	1–1	past	165.2	What color was the dress that she wore at her birthday lunch?
NUMERIC	1–1	past	164.5	How many Oscars has she won?
MEASURE	1–1	present	167.2	How high is the Millennium Wheel?
TIME	1–1	future	192.5	What date will this cease-fire begin?
LOCATION	2–0	past	172.1	The WTO has held meetings in what countries?
ADDRESS	1–1	present	143.5	What is the zip code of the American Enterprise Institute?
NAME	1–1	present	172.3	What is Ben's last name?
LANGUAGE	1–1	present	192.1	What does the acronym ETA stand for?
CREATION	2–0	past	164.2	What movies did she play in?
INSTANCE	5–5	past	214.7	Who were the five finalists in the pageant?
KIND	1–1	past	104.2	What type of vehicle dominated the show?
PART	1–1	present	162.1	Myeloma is cancer in what part of the body?

malisation rules are applied in a post-processing step, thereby dealing with active-passive alternations, inferred semantic information, normalisation of temporal expressions, and the disambiguation of noun-noun compounds. The resulting DRS is enriched with information about the original surface word-forms and parts of speech.

2.6 Answer Extraction

Given the DRS of the question (the Q-DRS), and a set of DRSs of the retrieved documents (the A-DRSs), each A-DRS is matched with the Q-DRS to find a potential answer. This process proceeds as follows: if the A-DRS contains a discourse referent of the answer type (see Section 2.2) matching will commence attempting to identify the semantic structure in the Q-DRS with that of the A-DRS. The result is a score between 0 and 1 indicating the amount of semantic material that could be matched. The background knowledge (such as hyponyms from WordNet) generated by the Question Analysis (see Section 2.3) is used to assist in the matching.

2.7 Answer Selection

The Answer Extraction component yields a list of answers and a matching score. Answers that are semantically identical are grouped together. This gives a new list of answers, ranked on matching score and frequency. Two methods of reranking were employed at the TREC-2006 exercise:

1. simple: sort on matching score, use highest frequency as tie-breaker;

2. combined: rank on a combination of matching score and frequency, assigning a weight of 0.9 to the matching score, and 0.1 to frequency.

The answer cardinality (see Section 2.2) determines the number of answers that are generated by the system, with a maximum of 10 answers if the upper bound of the answer cardinality is unspecified. Following (Dalmas and Webber, 2006), for some answer types (in particular TIME), answers that entail each other are identified and the answer with the most informative surface structure is ranked highest.

2.8 Processing *other*-questions

Since *other*-questions do not appear as ordinarily formulated questions, but the QA system expects questions phrased in English as input, they are automatically transformed into definition questions. This is simply done by substituting the target for the empty slot in "What is _ ?" and assigning it the answer type DEFINITION with answer cardinality 1–0. The answer extraction component deals with definition questions by finding sentences with the target as agent of an event. A higher matching score is given to sentence that contain superlative or temporal expressions.

3 Evaluation

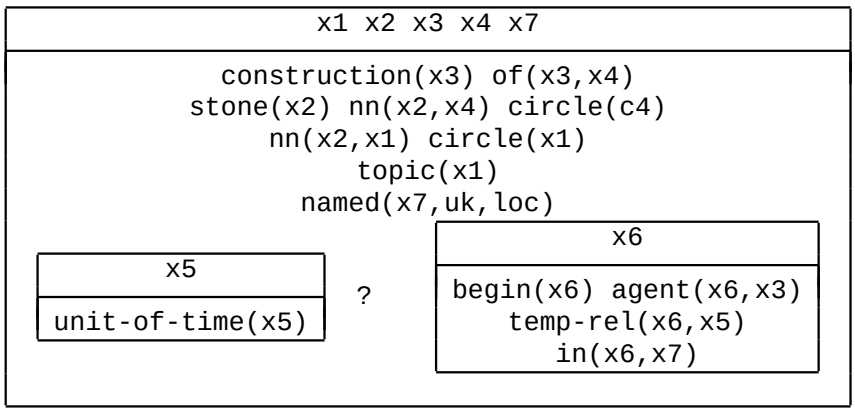
3.1 Experimental Setup

Three runs were submitted, all with different parameters with respect to the treatment of factoids, list, and other-questions. The parameters were the type of query: basic or complex (see Section 2.4), and reranking method: simple or combined (see Section 2.7). Table 2 summarises the runs and the parameters used.

This setup would tell us the following: (1) if Run 1 achieved higher results than Run 2, then the complex query method outperformed the basic one; (2) if Run 2 achieved higher results than Run 3, then the complex ranking methods would be better than the simple one.

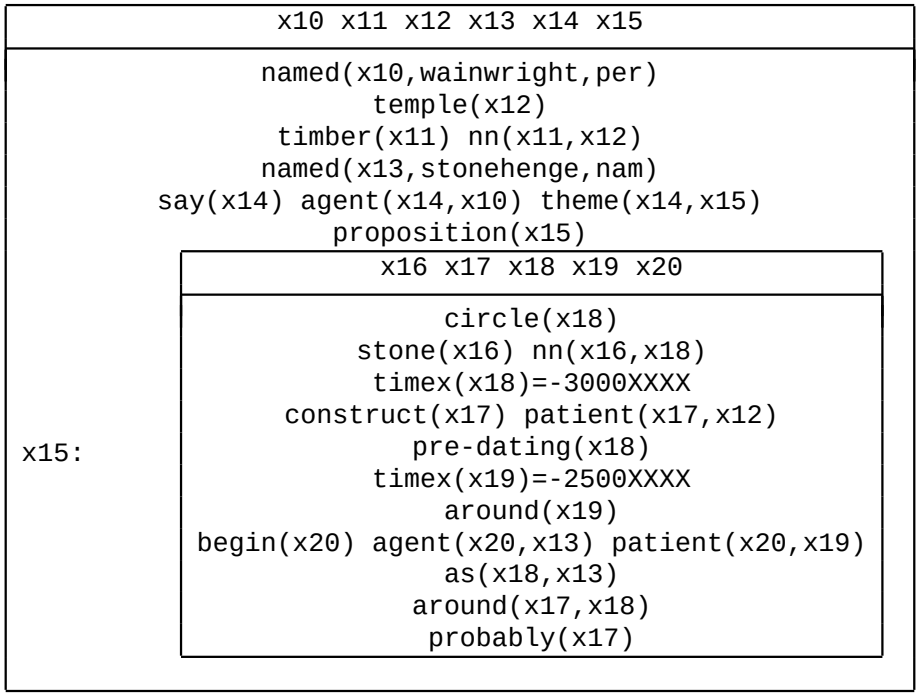
Question:	When did the construction of stone circles begin in the UK?
Target:	stone circles
ID:	169.1

Q-DRS:



Answer Type:	[tim:any]
Answer Tense:	past
Cardinality:	1–1
Query	#filreq(UK #weight(1 UK 4 stone 4 construction 4 circle 3 begin))
Context:	[XIE19971111.0069] Wainwright said that the timber temples were probably constructed around 3,000 BC, pre-dating stone circles, such as Stonehenge, which began around **2,500 BC**.

A-DRS:



Answer:	169.1 Roma2006run3 XIE19971111.0069 2,500 BC

Figure 2: System input and output for the factoid question 169.1 at TREC-2006.

Table 2: Description of the three runs at TREC-2006.

Run	Query Type	Reranking
Roma2006run1	complex	simple
Roma2006run2	basic	combined
Roma2006run3	basic	simple

3.2 TREC-2006 Judgements

Factoid questions formed the majority of the questions at the TREC-2006 QA evaluation exercise. The results of the La Sapienza system over 403 factoid questions are listed in Table 3 below, where U is the number of unsupported (correct but without a supporting document), X the number of inexact, L the number of locally correct (a later document in the Acquint corpus contradicts the answer), and R the number of correct answers.

Table 3: Results for *factoid*-questions, TREC-2006.

Run	U	X	L	R	Acc.	L.Acc.
Run 1	7	15	4	62	0.15	0.22
Run 2	16	15	4	68	0.17	0.26
Run 3	11	17	4	73	0.18	0.26
all	757	1163	151	4476	0.19	0.28

The last two columns of Table 3 show the accuracy (calculated on the basis of R) and lenient accuracy (calculated on the basis of U+X+L+R). In addition, it shows the summed scores of all participating systems at TREC-2006, a total of 59 runs. As the figures of accuracy show, the La Sapienza system performed slightly under the averaged accuracy. This was slighly disappointing, nonetheless the third run of the La Sapienza system had one correct answer that no other participating system managed to find—this was the answer to 169.1, as shown in Figure 2.

Table 4: Results (average F-scores) for *list* and *other*-questions, and per-series scores at TREC-2006.

Run	List	Other	Series
Run 1	0.12	0.14	0.14
Run 2	0.11	0.15	0.15
Run 3	0.13	0.16	0.16
median	0.09	0.13	0.13
best	0.43	0.25	0.39

There were 89 *list*-questions in total. These are evaluated by calculating the precision and recall for each question and then averaging their corresponding F-scores. The La Sapienza system achieved an average F-score higher than the median of all participating systems (Table 4). The results of the *other*-questions were encouraging, too: despite the fact that we didn't do anything sophisticated for dealing with other-questions, the obtained results were higher then the median of all 59 runs at TREC-2006. Also the per-series results were higher than the medium score of all participating systems.

Since for all types of questions Run 3 achieved the highest results, it can be concluded that the attempt to construct good queries failed, as it is outperformed by the baseline method, just using the target as query. Also the attempt on another reranking method, other than just using the question-answer matching score, but taking the frequency into account, didn't give better results.

Table 5: Distribution of answer-types on the TREC-2006 test set, for both *factoid*- and *list*-questions.

Answer-Type	Factoid	List	Total
INSTANCE	104	52	156
LOCATION	64	20	84
TIME	80		80
NUMERIC	67		67
MEASURE	40		40
CREATION	17	14	31
KIND	11	3	14
NAME	11		11
LANGUAGE	4		4
DESCRIPTION	2		2
PART	1		1
ATTRIBUTE	1		1
ADDRESS	1		1
	403	89	492

3.3 General Evaluation

The question analysis component of the La Sapienza system was evaluated by calculating the accuracy of answer type determination. Table 5 shows the distribution of all questions of the TREC-2006 test set over the inventory of answer types.

The overall accuracy was 82% over 492 questions (both *factoid* and *list*-questions). This was lower then expected. The low score is partly due to failed or incorrect parses, and partly due to the lack of appropriate rules that determine the answer type. Table 6 lists the determination accuracy for the different types. For some frequent types, such as MEASURE, KIND and NAME, this is rather low.

Finally, the performance of the La Sapienza QA system with respect to different answer-types was investigated. If the system performs better or worse for some answer-types, then this could give an indication where

Table 6: Wrongly assigned and accuracy of answer type determination, for all TREC-2006 questions.

Answer-Type	Number	Wrong	Accuracy
INSTANCE	156	21	0.87
LOCATION	84	12	0.86
TIME	80	4	0.95
NUMERIC	67	10	0.85
MEASURE	40	18	0.55
CREATION	31	6	0.81
KIND	14	10	0.29
NAME	11	6	0.45
LANGUAGE	4	1	0.75
DESCRIPTION	2	0	1.00
PART	1	0	1.00
ATTRIBUTE	1	0	1.00
ADDRESS	1	1	0.00
	492	89	0.82

the weaknesses or strong points of the overall system are. Table 7 gives an impression.

Table 7: Number of unsupported (U), inexact (X), locally correct (L), and correct (R) answers distributed over answer types, together with the achieved accuracy and lenient accuracy, at TREC-2006.

Answer-Type	U	X	L	R	Acc.	L.Acc.
INSTANCE	6	2	2	15	0.14	0.24
LOCATION	2	6		16	0.25	0.38
TIME	2	7	1	23	0.29	0.41
NUMERIC	1		1	8	0.12	0.15
MEASURE				2	0.05	0.05
CREATION				4	0.24	0.24
KIND					0.00	0.00
NAME		2		2	0.18	0.36
LANGUAGE				2	0.50	0.50
DESCRIPTION				1	0.50	0.50
	11	17	4	73	0.18	0.26

As Table 7 clearly shows, the La Sapienza system performs reasonably well on TIME and LOCATION questions. A likely explanation is that for these answer types, the named entities (time expressions and locations) are easier to find in texts than those corresponding to persons, organisations, or creative works. The system scored relatively low on answers of type NUMERIC and MEASURE, which is partly due to unsolved problems in answer typing.

Compared to what other systems are capable of (see Table 8), the La Sapienza system is particularly good at question with answer types TIME, CREATION, NAME, LANGUAGE, and DESCRIPTION. This table also demonstrates that many of the QA systems at TREC-2006 had difficulties with questions with the common answer types such as NUMERIC, MEASURE, CREATION, KIND, NAME DESCRIPTION, PART, and ATTRIBUTE.

Table 8: The average accuracy scores distributed over answer-types, for all 59 runs at TREC-2006.

Answer-Type	Factoid	Correct	Accuracy
INSTANCE	104	1255	0.20
LOCATION	64	1007	0.27
TIME	80	1007	0.21
NUMERIC	67	590	0.15
MEASURE	40	267	0.11
CREATION	17	114	0.11
KIND	11	80	0.12
NAME	11	54	0.08
LANGUAGE	4	52	0.22
DESCRIPTION	2	20	0.17
PART	1	4	0.07
ATTRIBUTE	1	9	0.15
ADDRESS	1	17	0.28

4 Conclusion

The La Sapienza QA system is based on a deep linguistic analysis of question and potential answers contexts and uses semantics to narrow down the number of answer candidates. Compared to other QA systems at TREC-2006, it performed slightly under par for *factoid*-questions, but better than average for *list* and *other*-questions.

The weak points of the La Sapienza system is the document retrieval (estimated loss of 20% of answers) and robust question analysis answer typing (looked particularly hard for some questions at this TREC) which probably caused another loss of around 20%.

The strong point of the system is that it performs really well on certain types of question, which probablt can be attributed to the ability of recognising the required type of named entities with high precision. A case in point are questions asking for temporal or locative expressions, for which the La Sapienza reaches relatively high accuracy scores.

Acknowledgements

Johan Bos is supported by a “Rientro dei Cervelli” grant (Italian Ministry for Research). He would like to thank Kisuh Ahn, Stephen Clark, James Curran, Tiphaine Dalmas, Dave Kor, Jochen Leidner, Diego Molla, Roberto Navigli, Malvina Nissim, Mark Steedman, Paola Verlardi, and Bonnie Webber for their feedback and help.

References

[Ahn et al. 2004] Kisuh Ahn, Johan Bos, Stephen Clark, James R. Curran, Tiphaine Dalmas, Jochen L. Leidner, Matthew B. Smillie, and Bonnie Webber (2004): Question answering with qed and wee at trec-2004. In E.M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*, NIST Special Publication 500-261, Gaithersburg, MD.

[Ahn et al. 2005] Kisuh Ahn, Johan Bos, James R. Curran, Dave Kor, Malvina Nissim, and Bonnie Webber (2005): Question answering with qed at trec-2005. In E.M. Voorhees and Lori P. Buckland, editors, *The Fourteenth Text Retrieval Conference (TREC 2005)*, NIST Special Publication 500-266, Gaithersburg, MD.

[Bos et al. 2004] J. Bos, S. Clark, M. Steedman, J.R. Curran, and Hockenmaier J. (2004): Wide-Coverage Semantic Representations from a CCG Parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, Geneva, Switzerland.

[Bos 2005] Johan Bos (2005): Towards wide-coverage semantic interpretation. In *Proceedings of Sixth International Workshop on Computational Semantics IWCS-6*, pages 42–53.

[Clark and Curran 2004] S. Clark and J.R. Curran (2004): Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL '04)*, Barcelona, Spain.

[Curran and Clark 2003] James R. Curran and Stephen Clark (2003): Language independent NER using a maximum entropy tagger. In *Proceedings of the Seventh Conference on Natural Language Learning (CoNLL-03)*, pages 164–167, Edmonton, Canada.

[Dalmas and Webber 2006] T. Dalmas and B. Webber (2006): Answer Comparison in Automated Question Answering. *Questions and Answers: Theoretical and Applied Perspectives. A Special Issue of the Journal of Applied Logic.*

[Fellbaum 1998] C. Fellbaum, editor (1998): *WordNet. An Electronic Lexical Database* The MIT Press.

[Hearst 1992] Marti A. Hearst (1992): Automatic aquisition of hyponyms from large text corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.

[Kamp and Reyle 1993] H. Kamp and U. Reyle (1993): *From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT* Kluwer, Dordrecht.

[Leidner et al. 2003] Jochen L. Leidner, Johan Bos, Tiphaine Dalmas, James R. Curran, Stephen Clark, Colin J. Bannard, Mark Steedman, and Bonnie Webber (2003): The QED open-domain answer retrieval system for TREC 2003. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, NIST Special Publication 500-255, pages 595–599, Gaithersburg, MD.

[Metzler and Croft 2004] D. Metzler and W.B. Croft (2004): Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, 40(5):735–750.

[Meyers et al. 1998] Adam Meyers, Catherine Macleod, Roman Yangarber, Ralph Grishman, Leslie Barrett, and Ruth Reeves (1998): Using nomlex to produce nominalization patterns for information extraction. In *Coling-ACL98 workshop Proceedings: the Computational Treatment of Nominals*, August.