Question Answering with SEMEX at TREC 2005

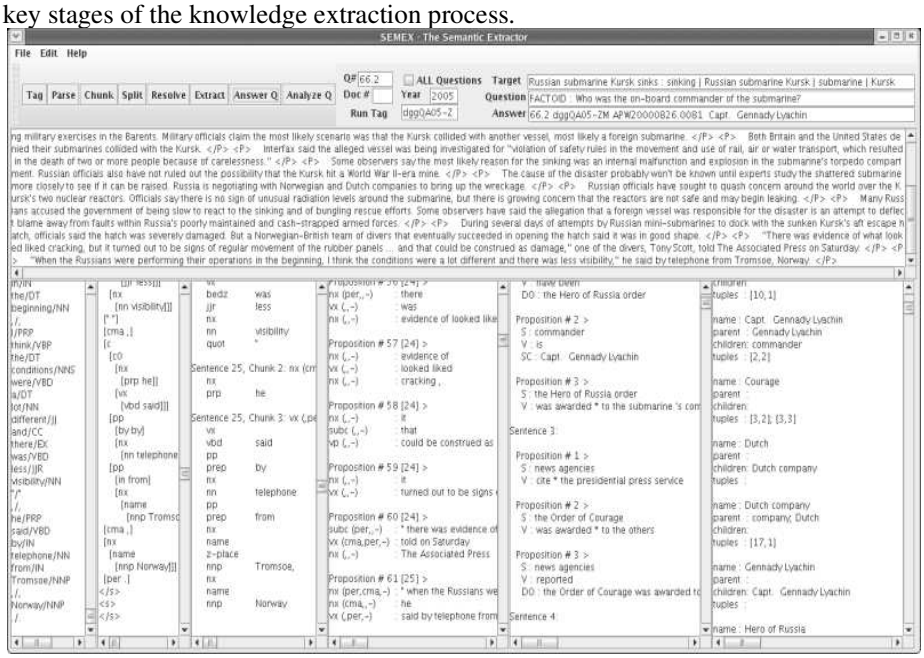
Demetrios G. Glinos
glinosd@saic.com

School of Computer Science
University of Central Florida
Orlando, FL 32816-2362

Abstract. We describe the SEMEX question-answering system and report its performance in the TREC 2005 Question Answering track. Since this was SEMEX's first year participating in the TREC evaluations, implementation teething pains were expected and indeed encountered. Nevertheless, performance against difficult factoid and list questions was supportive of the question answering approach that was implemented.

1 System Description

Our SEMEX (SEMantic EXtractor) tool is a test bed environment for evaluating and refining semantic extraction and question answering algorithms. SEMEX provides the graphical user interface shown in Figure 1 for viewing the intermediate results at key stages of the knowledge extraction process.



As shown in the figure, the document being processed appears in the horizontal text area at the top. The six vertically-oriented text areas below it display the intermediate results after the following key stages of the semantic extraction process:

1. Part of speech tagging
2. Partial parsing
3. Chunking
4. Sentence splitting
5. Resolution
6. Concept extraction

For the tagging component, SEMEX uses the Brill tagger [2], whose output is corrected for common tagger errors. Parsing is performed using Abney's Cass partial parser. SEMEX then applies a comprehensive set of empirically derived heuristics to build up phrases at the chunking stage. The resultant parse trees are then simplified and reduced to atomic propositions in the sentence splitting stage. Syntactic roles are assigned to the propositions and pronomial references are then resolved. And finally, concepts are extracted for each discourse entity identified in the resolved propositions.

SEMEX is presently configured to implement concept nodes that link to the resolved propositions in which they appear. The resolved propositions are represented as vectors whose components correspond to the key syntactic roles:

< subject, verb, gerund/infinitive, adverbials, indirect_object, direct_object >

The concept nodes are further organized into a hierarchy of “is-a” relationships that are derived from both the proposition set and the concept phrases themselves. Thus, a proposition for “space shuttle Discovery” will have a parent link to “space shuttle” which, in turn, will link to “shuttle.”

SEMEX provides text fields at the top of the GUI for entering the target and question, as shown in Figure 1. The figure also shows components for question number, year, run tag, and all question selection, which are important for TREC batch mode execution, but are not used for ad hoc processing.

As presently configured, SEMEX processes a question first by resolving any pronouns using the target, and then by tagging and parsing the question to produce a question vector or boolean combination of vectors of the same form shown above for propositions, except that the expected answer is replaced with a variable. The text field below the question shows the question vectors when the “analyze” button is pressed. The same field displays the answer when the “answer” button is pressed. To produce an answer, SEMEX performs a unification of the question vector or vectors with the relevant vectors retrieved from the concept hierarchy for the particular question. WordNet [3] was used in the unification process to improve recall.

2 Experiment Setup

SEMEX was modified to include components and logic necessary for executing in batch mode, so that it could process all questions for all targets in a single pass without human intervention. And since SEMEX did not possess an IR component, it was configured to make use of the "Top 50" document lists furnished by NIST for each target.

For factoid questions, SEMEX was configured to process top documents successively until a first answer was obtained. Once an answer was found for a factoid question, no attempt was made to search for a "better" answer. By contrast, for list questions, all readable documents were processed, and all answers obtained from any of them were reported. And for "other" questions, SEMEX reported the phrases for all propositions considered relevant in the database.

Two runs were submitted for official evaluation. They differed only in the input data file noise filtering. Although this did not result in significant change in the results, wall clock execution time for the run was reduced from 18 hours to 12.5 hours.

3 Experimental Results

SEMEX did not perform well in its first TREC evaluation. An examination of the results revealed a bug in the unification logic for adverbials that prevented finding correct answers for any question other than "when" questions. Also, when viewing in SEMEX the results of individual questions executed against single documents, it was evident that parsing performance was quite poor against the complex and often run-on sentences that characterize the newswire document collection. Parsing performance was particularly poor for documents that did not contain complete sentences, such as sports score articles, and documents that were collections of snippets from many sources, as the latter tended to have embedded parenthesized metadata and spurious embedded HTML codes.

Nevertheless, salient components of the official results for the two runs are summarized in Table 1, where Run 1 represents the run with the least noise filtering.

Score Category	***Criterion***	***Run 1 result***	***Run 2 result***
FACTOID	Accuracy	0.036	0.041
	NIL Precision	0.040	0.045
	NIL Recall	0.647	0.706
	# right + # inexact	14	18
LIST	Average F	0.005	0.005
OTHER	Average F	0.140	0.141
	# with recall >= .5	20	21
PER-SERIES	Average F	0.054	0.057
	# with F = 0	30	32

Table 1. SEMEX QA official results

These results show a nearly 29% increase in the number of correct plus inexact responses for factoid questions, confirming the value of reducing noise (primarily spurious HTML tags and metadata) in the input documents. The results also show that parsing performance was so poor against the documents for approximately 40% of the targets that they received per-series scores of zero.

Nevertheless, the strategy of reporting all relevant propositions in response to "other" questions seemed to find support even where parsing was poor, as over 25% of the targets had recall at least 50%.

In order to examine more deeply the performance of the question answering algorithms, the aforementioned bug permitting only "when" adverbials to find matches was corrected, as were a number of other small bugs related to algorithm scope and complexity. A complete run was executed, which took only approximately four hours wall clock time.

The factoid results for this informal run were hand graded using the NIST-furnished answer fact pattern files as a guide. However, grading was not as strict as the fact patterns would indicate. Thus, for example, we accepted as correct the answer "in the Barents Sea" for question 66.6, even though the fact pattern listed only "Barents Sea." Similarly, we accepted "Miss India Lara Dutta" for 67.1, compared with the listed pattern "Lara Dutta". Scored in this manner, the informal run achieved a factoid accuracy score of 26/362 = .072. Although this is not in absolute terms very high, which confirmed continued poor parsing performance, it nonetheless represented a 70% improvement over the previous runs, all in one-third the running time. List and "other" results were analyzed.

What is more revealing is that where acceptable parses were obtained, SEMEX appeared to operate well. For example, for the target *"Russian submarine Kursk sinks"* TREC question 66.7 asked the list question, *"Which U.S. submarines were reportedly in the area?"* For this question, SEMEX generated the question pattern:

<or> [S:*which][V:were][GI:][M:reportedly;in the area][IO:][DO:]
<and> [S:*ans][V:is][GI:][M:][IO:][DO:U.S. submarine]

and returned the answer "the Toledo". What is significant about this result is that the document in which this answer was found consisted of three sentences, the first two of which were:

<P> The second U.S. submarine in the Barents Sea when the Kursk sank was the Toledo, a Russian news agency reported Thursday. </P>
<P> The agency, Interfax, said the Toledo was in the area along with another U.S. submarine, the Memphis, during the Russian naval exercises in mid-August, when the Kursk sank, with the loss of 118 lives. </P>

In this example, the fact that the Toledo was a U.S. submarine is established in the first sentence of the document, while the fact that it was reportedly in the area when the Kursk sank was furnished by the second sentence. The connection between these two facts was provided by the concept hierarchy generated by SEMEX, in which the "Toledo" concept had the parent, "second U.S. submarine in the Barents Sea when

the Kursk", which in turn had the parent "second U.S. submarine" which, by a further link had the parent, "U.S. submarine", which was the desired class. We note that SEMEX was unable to find the second valid answer, "the Memphis", since SEMEX did not understand that vessel to have been in the area, although it did recognize it as a U.S. submarine.

Against the same target, SEMEX was able to find the correct answer, "Capt. Gennady Lyachin," in response to question 66.2, a result which only 2 out of 71 total QA runs were able to achieve. The factoid question involved was *"Who was the onboard commander of the submarine?"* The relevant input sentence was *"The Hero of Russia order, one of the country's highest honors, was awarded to the submarine's commander, Capt. Gennady Lyachin."* SEMEX was able to find the answer in this case because (a) the SEMEX sentence splitting logic had split off the apposition and had created the copular sentence that "commander" "is" "Capt. Gennady Lyachin", and (b) SEMEX was able to associate "the submarine's commander" with "the onboard commander of the submarine."

In a similar vein, question 120.3 ask *"What organization did she found?"* for the target *"Rose Crumb."* For this question, SEMEX returned the correct answer: *"the volunteer Hospice of Clallam County,"* which only 7 other runs were able to answer correctly. Examining SEMEX operation in this case revealed that the relevant input sentence was:

> *"Crumb, who founded the volunteer Hospice of Clallam County, Wash., in 1978, said she has learned ``courage, patience and acceptance'' by working with families to provide their dying relative with a dignified end."*

SEMEX was able to split off the relative clause and create a separate proposition for it. That propostion, as shown in SEMEX's resolution window, was:

> Proposition # 5 >
> S : Crumb
> V : founded
> DO : the volunteer Hospice of Clallam County

which was easily matched against the question because "Hospice of Clallam County" was recognized as a business organization.

On another topic, Question 87.3 asked the factoid question, *"What Nobel Prize was Fermi awarded in 1938?"* For this question, SEMEX correctly returned *"the Nobel Prize for Physics"*, which only 10 other runs achieved. This was made possible by the parent-child relationship present in the concept hierarchy between "Nobel Prize" and the answer returned. Indeed, a separate test has shown that SEMEX would still have returned the same result if the question had asked what "prize" was Fermi awarded, again as a consequence of the concept hierarchy, and even if the question had asked about an "award", which though not in the hierarchy was nevertheless a WordNet hypernym of "Nobel Prize."

In reviewing SEMEX's performance in the informal run, we found that there were questions which WordNet would consider inaccurately stated. For example, for the target *"Miss Universe 2000 crowned,"* factoid question 67.4 asked, *"Where was the*

contest held?" For this question, SEMEX was unable to find an answer, even though it had correctly split off the following proposition:

Proposition # 2 >
S : Miss India Lara Dutta
V : was crowned * during the 49th Miss Universe pageant
* in Nicosia, Cyprus * early Saturday
DO : Miss Universe 2000

Examining SEMEX's internal operation for this question revealed that the reason the location adverbial *"in Nicosia, Cyprus"* was not matched is because a "pageant" does not have "contest" in any of its WordNet synsets.

4. Conclusions

Considering as a whole the experimental results reviewed above, we are drawn to the following observations:

1. For our question answering approach, parsing performance is crucial.
2. Spurious HTML tags, parenthetical metadata, and other non-sentential content in text documents can degrade performance significantly.
3. Splitting sentences into distinct propositions prior to concept extraction is beneficial.
4. Where adequate parsing performance was obtained, SEMEX's ability to find answers in complex situations has demonstrated the value of a concept hierarchy encoding parent-child concept relationships and the robustness that may be achieved through the use of WordNet.

Based on these observations, we conclude that the question answering approach we have implemented bears further study.

References

1. Abney, S.: Partial Parsing via Finite-State Cascades. In: Proceedings of Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information. Prague, Czech Republic (1996) 8-15.
2. Brill, E.: Some Advances in Part of Speech Tagging. In: Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94). Seattle, Washington (1994)
3. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge London (1998)