Syntactic Clues and Lexical Resources in Question-Answering

Kenneth C. Litkowski
CL Research
9208 Gue Road
Damascus, MD 20872
ken@clres.com

Abstract

CL Research's question-answering system (DIMAP-QA) for TREC-9 significantly extends its semantic relation triple (logical form) technology in which documents are fully parsed and databases built around discourse entities. This extension further exploits parsing output, most notably appositives and relative clauses, which are quite useful for question-answering. Further, DIMAP-QA integrated machine-readable lexical resources: a full-sized dictionary and a thesaurus with entries linked to specific dictionary definitions. The dictionary's 270,000 definitions were fully parsed and semantic relations extracted to provide a MindNet-like semantic network; the thesaurus was reorganized into a WordNet file structure. DIMAP-QA uses these lexical resources, along with other methods, to support a just-in-time design that eliminates preprocessing for named-entity extraction, statistical subcategorization patterning, anaphora resolution, ontology development, and unguided query expansion. (All of these techniques are implicit in DIMAP-QA.)

The best official scores for TREC-9 are 0.296 for sentences and 0.135 for short answers, based on processing 20 of the top 50 documents provided by NIST, 0.054 and 0.083 below the TREC-9 averages. The initial post-hoc analysis suggests a more accurate assessment of DIMAP-QA's performance in identifying answers is 0.485 and 0.196. This analysis also suggests that many failures can be dealt with relatively straightforwardly, as was done in improving performance for TREC-8 answers to 0.803 and 0.597 for sentences and short answers, respectively.

1. Introduction

TREC-9 DIMAP-QA proceeded from last year's version by first removing many shortcomings noted there (where it was suggested that the official 250-byte, or sentence, score of 0.281 could be raised to an estimated 0.482) by including documents not processed, resolving parsing problems affecting both questions and documents, and resolving triple extraction problems. Dealing with these problems improved the score to 0.550. DIMAP-QA was then extended to extract 50-byte answers, with feedback to the sentence extraction (i.e., when a viable short answer was recognized, its sentence was given a higher score). This extension focused on developing question-specific routines for extracting short answers based on the discourse entities and the types of semantic relations in which they participated. This improved scores to 0.740 for sentences and 0.493 for short answers, suggesting that a substantial portion of question-answering can be achieved without special pre-processing. At this point in development, the lexical resources were integrated. Although these resources could have been used directly to answer questions, the just-in-time model used them instead for substantiation. For example, in "where" questions, definitions provided a background set of discourse entities used in evaluating document sentences. For "what" questions (e.g., "what country"), dictionary definitions were examined to determine whether a document discourse entity was defined as or had the hypernym "country". If no match, the thesaurus was examined to determine if the hypernym for a document discourse entity was in the same thesaurus category (e.g., as "country" where "Belgium" is defined as a "kingdom"). Incorporation of these lexical resources improved the TREC-8 scores to 0.803 for sentences and 0.597 for short answers.

DIMAP-QA is a part of the DIMAP dictionary creation and maintenance software, which is primarily designed for making machine-readable dictionaries machine-tractable and suitable for NLP tasks, with some components intended for use as a lexicographer's workstation.[1] The TREC-9 QA track provided an opportunity for experimenting with the limits of

[1]DIMAP, including the question-answering component, is available from CL Research. Demonstration and experimental versions are available at http://www.clres.com.

question-answering based only on syntactical clues and for examining use of computational lexical resources (dictionary and thesaurus). The development of the system for TREC-9 and the analysis of failures provides a good delineation of the limits of different types of evidence and the role of lexical resources.

2. Problem Description

Participants in the TREC-9 QA track were provided with 693 unseen questions to be answered from the TREC CD-ROMs, (about 1 gigabyte of compressed data), containing documents from the *Foreign Broadcast Information Service*, *Los Angeles Times*, *Financial Times*, *Wall Street Journal*, *Associated Press Newswire*, and *San Jose Mercury News*. These documents were stored with SGML formatting tags. Participants were given the option of using their own search engine or of using the results of a "generic" search engine. CL Research chose the latter, relying on the top 50 documents retrieved by the search engine. These top documents were provided simultaneously with the questions.

Participants were then required to answer the 693 questions in either 50-byte answers or by providing a sentence or 250-byte string in which the answer was embedded. For each question, participants were to provide 5 answers, with a score attached to each for use in evaluating ties.[2] NIST evaluators then judged whether each answer contained a correct answer. Scores were assigned as the inverse rank. If question q contained a correct answer in rank r, the score received for that answer was 1/r. If none of the 5 submissions contained a correct answer, the score received was 0. The final score was then computed as the average score over the entire set of questions.

CL Research submitted 4 runs, 2 each for the 250- and 50-byte restrictions, one analyzing only the top 10 documents and the other only the top 20 documents, to examine whether performance was degraded in going from 10 to 20 documents.

3. System Description

The CL Research prototype system consists of four major components: (1) a sentence splitter that separated the source documents into individual sentences; (2) a parser which took each sentence and parsed it, resulting in a parse tree containing the constituents of the sentence; (3) a parse tree analyzer that identified important elements of the sentence and created semantic relation triples stored in a database; and (4) a question-answering program that (a) parsed the question into the same structure for the documents, except with an unbound variable, and (b) matched the question database records with the document database to answer the question. The matching process first identified candidate sentences from the database, extracted short answers from each sentence, developed a score for each sentence, and chose the top 5 sentences (and their short answers) for submission.

3.1 Sentence Identification in Documents

The parser (described more fully in the next section) contains a function to recognize sentence breaks. However, the source documents do not contain crisply drawn paragraphs that could be submitted to this function. Thus, a sentence could be split across several lines in the source document, perhaps with intervening blank lines and SGML formatting codes. As a result, it was first necessary to reconstruct the sentences, interleaving the parser sentence recognizer.

At this stage, we also extracted the document identifier and the document date. Other SGML-tagged fields were not used. The question number, document number, and sentence number provided the unique identifier when questions were answered.

TREC-9 added 3 document collections (*Wall Street Journal*, *Associated Press Newswire*, and *San Jose Mercury News*). Although we had tested processing of these document types before the test suite was made available, we had not captured nuances not described in the DTDs. As a result, there were many "bombs" that occurred in processing the top documents; many of the problems had to be fixed during the final processing. Although this violates the strict rule against making changes after the questions are made available, these changes did not go to the heart of the question-answering, but only to the ability of the system to process the documents. After submission, further nuances affecting system performance were identified, most notably in the omission of important textual material ("lead paragraphs") in the *Wall Street Journal* and the *San Jose Mercury News* and the combining of multiple sentences from *Associated Press* documents (because of the way quoted material was handled). These

[2]Although this statement appears in one of the problem specifications, the score is not used and only the position of the answer is considered.

problems had an effect on performance, as described below.

For the TREC-9 QA runs submitted to NIST, the top 20 documents (as ranked by the search engine) were analyzed (30 were processed for 250 of the questions). Overall, this resulted in processing 14605 documents (up from 1977 in TREC-8) from which 422,562 (up from 63,118) sentences were identified and presented to the parser. Thus, we used an average of 28.9 (down from 31.9) sentences per document or 290 sentences for the 10-document set, 580 for the 20-document set, and 870 for the 30-document set for each question.

3.2 Parser

The parser in DIMAP (provided by Proximity Technology, Inc.) is a grammar checker that uses a context-sensitive, augmented transition network grammar of 350 rules, each consisting of a start state, a condition to be satisfied (either a non-terminal or a lexical category), and an end state. Satisfying a condition may result in an annotation (such as number and case) being added to the growing parse tree. Nodes (and possibly further annotations, such as potential attachment points for prepositional phrases) are added to the parse tree when reaching some end states. The parser is accompanied by an extensible dictionary containing the parts of speech (and frequently other information) associated with each lexical entry. The dictionary information allows for the recognition of phrases (as single entities) and uses 36 different verb government patterns to create dynamic parsing goals and to recognize particles and idioms associated with the verbs (the context-sensitive portion of the parser).

The parser output consists of bracketed parse trees, with leaf nodes describing the part of speech and lexical entry for each sentence word. Annotations, such as number and tense information, may be included at any node. The parser does not always produce a correct parse, but is very robust since the parse tree is constructed bottom-up from the leaf nodes, making it possible to examine the local context of a word even when the parse is incorrect. In TREC-9, parsing exceptions occurred for only 69 sentences out of 422562 (0.0002, down from 0.008), with another 131 "sentences" (usually tabular data) not submitted to the parser. Usable output was available despite the fact that there was at least one word unknown to the parsing dictionary in 33,467 sentences (7.9 percent).

3.3 Document and Question Database Development

A key step of DIMAP-QA is analysis of the parse tree to extract semantic relation triples and populate the databases used to answer the question. A **semantic relation triple** consists of a discourse entity, a semantic relation which characterizes the entity's role in the sentence, and a governing word to which the entity stands in the semantic relation. A triple is generally equivalent to a logical form (where the operator is the semantic relation) or a conceptual graph, except that a semantic relation is not strictly required, with the driving force being the discourse entity.

The first step of discourse processing is identification of suitable discourse entities. For TREC-8, this involved analyzing the parse tree **node** to extract numbers, adjective sequences, possessives, leading noun sequences, ordinals, time phrases, predicative adjective phrases, conjuncts, and noun constituents as discourse entities. To a large extent, named entities, as traditionally viewed in information extraction, are identified as discourse entities (although not specifically identified as such in the databases). For TREC-9, the parse output was further mined, more fully exploiting the syntactic relations between sentence constituents. The most notable of these was the characterization of various forms of appositives (parenthesized expressions, relative clauses, and true appositives), which frequently provide the answers to questions.

The semantic relations in which entities participate are intended to capture the semantic roles of the entities, as generally understood in linguistics. This includes such roles as agent, theme, location, manner, modifier, purpose, and time. For TREC-9, we did not fully characterize the entities in these terms, but generally used surrogate place holders. These included "SUBJ," "OBJ", "TIME," "NUM," "ADJMOD," and the prepositions heading prepositional phrases. Appositive phrases were characterized by identifying the sentence word they modified and the beginning and ending words of the phrase; their use is described particularly for answering Who and What questions.

The governing word was generally the word in the sentence that the discourse entity stood in relation to. For "SUBJ," "OBJ," and "TIME," this was generally the main verb of the sentence. For prepositions, the governing word was generally the noun or verb that the prepositional phrase modified. (Because of the context-sensitive dynamic parsing goals that were added when a verb or a governing noun was recognized, it was possible to identify what was modified.) For the adjectives and numbers, the governing word was generally the noun that was modified.

The semantic relation and the governing word were not identified for all discourse entities, but a record for each entity was still added to the database for the sentence. Overall, 4,149,106 semantic relation triples were created (up from 467,889) in parsing the 422,562 sentences, an average of 9.8 triples per sentence (up from 7.4 in TREC-8).

The same functionality was used to create database records for the 693 questions. The same parse tree analysis was performed to create a set of records for each question. The only difference is that one semantic relation triple for the question contained an unbound variable as a discourse entity, corresponding to the type of question. The question database contained 2272 triples (for 693 questions), an average of 3.3 triples per question. This is down from 4.5 triples per question in TREC-8. This is indicative of the fact that the questions were "simpler", making them more difficult to answer, since there was less information on which to match.

3.4 Lexical Resources

A major addition to the question-answering system for TREC-9 QA was the integration of a machine-tractable dictionary and thesaurus. These were provided in machine-readable form by The Macquarie Library Pty Ltd of Australia. The dictionary, known as Big Mac, was converted into a format suitable for uploading into DIMAP dictionaries, during which most of the raw data were put into specific fields of a DIMAP dictionary (e.g., headword, part of speech, definitions, example usages, and many "features" characterizing syntactic properties and other information, particularly a link to Macquarie's thesaurus and identification of a "derivational" link for undefined words to their root form).

After conversion and upload, the entire dictionary of 270,000 definitions was parsed to populate the raw dictionary data by adding semantic relations links with other words. The most important result was the identification of the hypernyms of each sense. Other relations include synonyms (discernible in the definitions), typical subjects and objects for verbs, and various semantic components (such as manner, purpose, location, class membership, and class inclusion). This dictionary, accessed during the question-answering process, is thus similar in structure to MindNet (Richardson, 1997).

The Macquarie thesaurus was provided in the form of a list of the words belonging to 812 categories, which are broken down into paragraphs (3 or 4 for each part of speech) and subparagraphs, each containing about 10 words that are generally synonymous. We were also provided (Green, 2000) with a set of perl scripts for inverting the thesaurus data into alphabetical order, where each word or phrase was listed along with the number of entries for each part of speech, and an entry for each distinct sense identifying the category, paragraph, and subparagraph to which the word or phrase belongs.

The resultant thesaurus is thus in the precise format of the combined WordNet index and data files ((Fellbaum, 1998)), facilitating thesaurus lookup.

3.5 Question Answering Routines

For TREC-9, a database of documents was created for each question, as provided by the NIST generic search engine. A single database was created for the questions themselves. The question-answering consisted of matching the database records for an individual question against the database of documents for that question.

The question-answering phase consists of three main steps: (1) coarse filtering of the records in the database to select potential sentences, (2) detailed analysis of the question to set the stage for detailed analysis of the sentences according to the type of question, establishing an initial score of 1000 for each sentence, (3) extracting possible short answers from the sentences, with some adjustments to the score, based on matches between the question and sentence database records and the short answers that have been extracted and (4) making a final evaluation of the match between the question's key elements and the short answers to arrive at a final score for the sentence. The sentences and short answers were then ordered by decreasing score for creation of the answer files submitted to NIST.

3.5.1 Coarse Filtering of Sentences

The first step in the question-answering phase was the development of an initial set of sentences. The discourse entities in the question records were used to filter the records in the document database. Since a discourse entity in a record could be a multiword unit (MWU), the initial filtering used all the individual words in the MWU. Question and sentence discourse entities were reduced to their root form, eliminating issues of tense and number. All words were reduced to lowercase, so that issues of case did not come into play during this filtering step. Finally, it was not necessary for the discourse entity in the sentence database to have a whole word matching a string from the question database. Thus, in this step, all records were selected from the document database having a discourse entity that contained a substring that was a word in the question discourse entities.

MWUs were analyzed in some detail to determine their type and to separate them into meaningful named entities. We examined the capitalization pattern of a phrase and whether particular subphrases were present in the Macquarie dictionary. We identified phrases such as "Charles Lindbergh" as a person (and hence possibly referred to as "Lindbergh"), "President McKinley" as a person with a title (since "president" is an uncapitalized word in the Macquarie dictionary), "Triangle Shirtwaist fire" as a proper noun followed by a common noun (hence looking for either "Triangle Shirtwaist" or "fire" as discourse entities).

The join between the question and document databases produced an initial set of unique (document number, sentence number) pairs that were passed to the next step.

3.5.2 Identification of Key Question Elements

As indicated above, one record associated with each question contained an unbound variable as a discourse entity. The type of variable was identified when the question was parsed and this variable was used to determine which type of processing was to be performed.

The question-answering system categorized questions into six types (usually with typical question elements): (1) **time** questions ("when"), (2) **location** questions ("where"), (3) **who** questions ("who" or "whose"), (4) **what** questions ("what" or "which," used alone or as question determiners), (5) **size** questions ("how" followed by an adjective), and (6) **number** questions ("how many"). Other question types not included above (principally "why" questions or non-questions beginning with verbs "name the ...") were assigned to the **what** category, so that question elements would be present for each question.

Some adjustments to the questions were made. There was a phase of consolidating triples so that contiguous named entities were made into a single triple. Then, it was recognized that questions like "what was the year" or "what was the date" and "what was the number" were not **what** questions, but rather **time** or **number** questions. Questions containing the phrase "who was the author" were converted into "who wrote"; in those with "what is the name of", the triple for "name" was removed so that the words in the "of" phrase would be identified as the principal noun. Other phraseological variations of questions are likely and could be made at this stage.

Once the question type had been determined and the initial set of sentences selected, further processing took place based on the question type. Key elements of the question were determined for each question type, with some specific processing based on the particular question type. In general, we determined the key noun, the key verb, and any adjective modifier of the key noun for each question type. For **who** questions, we looked for a year restriction. For **where** questions, we looked up the key noun in the Macquarie dictionary and identified all proper nouns in all its definitions (hence available for comparison with short answers or other proper nouns in a sentence). For **what** questions, we looked for a year restriction, noted whether the answer could be the object of the key verb, and formed a base set of thesaurus categories for the key noun. For **size** questions, we identified the "size" word (e.g., "far" in "how far"). For **number** questions, we also looked for a year restriction.

3.5.3 Extraction of Short Answers

After the detailed question analysis, processing for each question then examined each selected sentence, attempting to find a viable short answer and giving scores for various characteristics of the sentence. For **time**, **location**, **size**, and **number** questions, it was possible that a given sentence contained no information of the relevant type. In such cases, it was possible that a given sentence could be completely eliminated. In general, however, a data structure for a possible answer was initialized to hold a 50-byte answer and the sentence was assigned an initial score of 1000. An initial adjustment to the score was given for each sentence by comparing the question discourse entities (including subphrases of MWUs) with the sentence discourse entities, giving points for their presence and additional points when the discourse entities stood in the same semantic relation and had the same governing word as in the question.

1. Time Questions - The first criterion applied to a sentence was whether it contained a record that has a TIME semantic relation. The parser labels prepositional phrases of time or other temporal expressions (e.g., "last Thursday"); database records for these expressions were given a TIME semantic relation. We also examined triples containing "in" or "on" as the governing word (looking for phrases like "on the 21st", which may not have been characterized as a TIME phrase) or numbers that could conceivably be years. After screening the database for such records, the discourse entity of such a record was then examined further. If the discourse entity contained an integer or any of its words were marked in the parser's dictionary as representing a time period, measurement time, month, or weekday, the discourse entity was selected as a potential answer.

2. Where Questions - Each sentence was examined for the presence of "in", "at", "on", "of", or "from" as a semantic relation, or the presence of a capitalized word (not present in the question) modifying the key noun. The discourse entity for that record was selected as a potential answer. Discourse entities from "of" triples were slightly disfavored and given a slight decrease in score. If the answer also occurred in a triple as a governing word with a HAS relation, the discourse entity from that triple was inserted into the answer as a genitive determiner of the answer.

3. Who Questions - The first step in examining each sentence looked for the presence of appositives, relative clauses, and parentheticals. If a sentence contained any of these, an array was initialized to record its modificand and span. The short answer was initialized to the key noun. Next, all triples of the sentence were examined. First, the discourse entity (possibly an MWU) was examined to determine the overlap between it and the question discourse entities. The number of hits was then added to all appositives which include the word position of the discourse entity within its span. (A sentence could have nested appositives, so the number of hits can be recorded in multiple appositives.)

The next set steps involved looking for triples whose governing word matched the key verb, particularly the copular "be" and the verb "write". For copular verbs, if the key noun appeared as the subject, the answer was the object, and vice versa. For other verbs, we looked for objects matching the key noun, then taking the subject of the verb as the answer. A test was included here for examining whether the key noun is in the definition, a hypernym, or thesaurus category of the discourse entity, but this was not tested and was removed when the system was frozen.

Another major test of each discourse entity that contained a substring matching the key noun was whether it was modified by an appositive. If this was the case, the appositive was taken as a possible short answer; the discourse entities of the appositive were then concatenated into a short answer. Numerical and time discourse entities were also examined when there was a date restriction specified in the question to ascertain if they could be years, and if so, whether they matched the year restriction. In the absence of a clear sentence year specification, the document date was used.

4. What Questions - The first step in examining the sentences was identical to that of the **who** questions, namely, looking for appositives in the sentence and determining whether a discourse entity had overlaps with question discourse entities. If the key noun was a part of a discourse entity, we would note the presence of the key noun; if this occurrence was in a discourse entity identified as an adjective modifier, the modificand was taken as a short answer and if this short answer was itself a substring of another sentence discourse entity, the fuller phrase was taken as the answer. Similarly, when the key noun was a proper part of a discourse entity and began the phrase (i.e., a noun-noun compound), the remaining part was taken as the short answer.

As with **who** questions, if the key noun was identified as the modificand of an appositive, the appositive was taken as the possible answer. Similarly to **who** questions, we also looked for the copular "be" with the key noun as either the subject or object, taking the other as a possible answer. When the key verb was "have" and the key noun was equal to the object, the subject of "have" was taken as the short answer. In cases like these, we would also insert any adjective modifiers of the noun discourse entities at the beginning of the short answer.

If the key noun was not equal to the discourse entity of the triple being examined, we tested whether the key noun against the DIMAP-enhanced Macquarie dictionary, looking for its presence (1) in the definition of the discourse entity, (2) as a hypernym of the discourse entity, or (3) in the same Macquarie thesaurus category. (For example, in examining "Belgium" in response to the question "what country", where country is not in definition and is not a hypernym, since it is defined as a "kingdom", we would find that "country" and "kingdom" are in the same thesaurus category.) Finally, as with **who** questions, we examined TIME and number discourse entities for the possible satisfaction of year restrictions.

5. Size Questions - For these questions, each triple of a selected sentence was examined for the presence of a NUM semantic relation or a discourse entity containing a digit. If a sentence contained no such triples, it was discarded from further processing. Each numerical discourse entity was taken as a possible short answer in the absence of further information. However, since a bare number was not a valid answer, we looked particularly for the presence of a measurement term associated with the number. This could be either a modificand of the number or part of the discourse entity itself, joined by a hyphen. If the discourse entity was a tightly joined number and measurement word or abbreviation (e.g., "6ft"), the measurement portion was separated out for lookup. The parsing dictionary characterizes measurement words as having a "measures", "unit", "MEASIZE", or "abbr" part of speech, so the modificand of the number was tested against these. If not so present in the parsing dictionary, the Macquarie definition was examined for the presence of the word "unit". When a measurement word was identified, it was concatenated with the number to provide the short answer.

6. Number Questions - The same criterion as used in size questions was applied to a sentence to see whether it contained a record that has a NUM semantic relation. If a selected sentence had no such triples, it was effectively discarded from further analysis. In sentences with NUM triples, the number itself (the discourse entity) was selected as the potential answer. Scores were differentially applied to these sentences so that those triples where the number modified a discourse entity equal to the key noun were given the highest number of points. TIME and NUM triples potentially satisfying year specifications were also examined to see whether a year restriction was met. In the absence of a clear sentence year specification, the document date was used.

3.5.4 Evaluation of Sentence and Short Answer Quality

After all triples of a sentence were examined, the quality of the sentences and short answers was further assessed. In general, for each question type, we assessed the sentence for the presence of the key noun, the key verb, and any adjective qualifiers of the key noun. The scores were increased significantly if these key items were present and decreased significantly if not. In the absence of a clear sentence year specification (for **who**, **what**, and **number** questions containing a year restriction), the document date was used. For certain question types, there were additional checks and possible changes to the short answers.

For **location** questions, where we accumulated a set of proper nouns found in the definition of the key noun, the score for a sentence was incremented for the presence of those words in the sentence. Proper nouns were also favored, and if two answers were found, a proper noun would replace a common noun; proper nouns also present as proper nouns in the Macquarie dictionary were given additional points. Similarly, if a sentence contained several prepositional phrases, answers from "in" phrases replaced those from "of" or "from" phrases. For questions in which the key verb was not "be", we tested the discourse entities of the sentence against the DIMAP-enhanced Macquarie dictionary to see whether they were derived from the key verb (e.g., "assassination" derived from "assassinate").

For **who** and **what** questions, when a sentence contained appositives and in which satisfactory short answers were not constructed, we examined the number of hits for all appositives. In general, we would construct a short answer from the modificand of the appositive with the greatest number of hits. However, if one appositive was nested inside another, and had the same number of hits, we would take the nested appositive. For these questions, we also gave preference to short answers that were capitalized; this distinguished short answers that were mixed in case.

For these two question types, we also performed an anaphora resolution if the short answer was a pronoun. In these cases, we worked backward from the current sentence until we found a possible proper noun referent. As we proceeded backwards, we also worked from the last triple of the each sentence. If we found a plausible referent, we used that discourse entity as the short answer and the sentence in which it occurred as the long answer, giving it the same score as the sentence in which we found the pronoun.

For **size** questions, we deprecated sentences in which we were unable to find a measurement word. We also looked for cases in which the discourse entities in several contiguous triples has not been properly combined (such as number containing commas and fractions), modifying the short answers in such cases.

After scores have been computed for all sentences submitted to this step, the sentences are sorted on decreasing score. Finally, the output is constructed in the desired format (for both 50-byte and 250-byte answers), with the original sentences retrieved from the documents. If a sentence is longer than 250 bytes, the string is reduced based on where the short answer appears in the sentence.

4. TREC-9 Q&A Results

CL Research submitted 4 runs, 2 each for the 50- and 250-byte lengths; the official scores for these runs are shown in Table 1. The score is the mean reciprocal rank of the best answer over all 682 questions that were included in the final judgments. The score of 0.287 for run clr00s1 means that, over all questions, the CL Research system provided a sentence with a correct answer as slightly better than 4th position. This compares to an average score of 0.350 among all submissions for the TREC-9 QA 250-byte answers (i.e., a correct answer slightly better than the 3rd position).

Table 1. CL Research Run Scores

Run	Doc. Num.	Type	Score	TREC Ave.
clr00s1	10	250-byte	0.287	0.350
clr00b1	10	50-byte	0.119	0.218
clr00s2	20	250-byte	0.296	0.350
clr00b2	20	50-byte	0.135	0.218

The CL Research runs differ in the number of documents of the top 50 documents provided by the generic search engine that were processed. As will be discussed below, the number of documents processed reflects a point of diminishing returns in finding answers from the top documents. Table 2 shows the number of questions for which answers were found at any rank for the 682 questions.

Table 2. Answers Found (682)

Run	Doc. Num.	Type	Num	Pct.
clr00s1	10	250-byte	289	0.424
clr00b1	10	50-byte	113	0.166
clr00s2	20	250-byte	296	0.434
clr00b2	20	50-byte	132	0.194

5. Analysis

DIMAP-QA added many components to the system used in TREC-8. The analysis that follows examines the failures of this year's system, along with a description of the incremental steps implemented in dealing with last year's failures. In this way, we hope to capture the characteristics of the question-answering process and the significance of specific components.

As mentioned above, we only processed the top 20 documents provided by NIST. Table 3 clearly indicates that, after the first 10 documents, the amount of incremental improvement from processing more documents is quite small. This table indicates that the CL Research results might better be interpreted in terms of the questions that could possibly have been answered. Table 4 makes these adjustments.

Table 3. Highest ranked top document containing strict answer string

Document Number	Number of Questions
1-10	474
11-20	38
21-30	21
31-40	18
41-50	12
None	130

Table 4. Adjusted scores for documents attempted

Run	Doc. Num.	Type	Score	Adj. Score
clr00s1	10	250-byte	0.287	0.412
clr00b1	10	50-byte	0.119	0.170
clr00s2	20	250-byte	0.296	0.394
clr00b2	20	50-byte	0.135	0.179

The significant difference between the unadjusted and adjusted scores raises an important question: is the question-answering track measuring retrieval performance or question-answering ability? It was noted earlier that the number of semantic relation triples for the questions had declined from 4.5 in TREC-8 to 3.3 in TREC-9. One of these triples contains a question element, so the decline in information content is about one-third. As a result, this year's questions, while being simpler to state, are actually more difficult to answer. This has meant that the likelihood of the retrieval system retrieving a relevant document much less.

While this makes it more difficult for systems relying on the NIST top documents, it also raises the question of what might be an appropriate retrieval strategy. CL Research experimented with the Macquarie dictionary in support of answers to **location** questions (the only "simple" questions in TREC-8, so this strategy was only implemented for that question type in TREC-9). While this strategy may help CL Research performance on other question types, it does not help the retrieval performance shown in Table 3. What it does suggest is that dictionary lookup can usefully be employed in rephrasing a question for retrieving relevant documents. Thus, for example, instead of retrieving birth announcements for "Who is Maria Theresa?", the retrieval engine can search for "archduchess of Austria, queen of Hungary and Bohemia" in addition to "Maria Theresa".

In making improvements to DIMAP-QA for TREC-9, we began by removing many shortcomings noted there (Litkowski, 2000). First, we included documents not processed. Next, we resolved several "bugs", parsing problems affecting both questions and documents and problems in the extraction of semantic relation triples. Dealing with these problems improved the score to 0.550, better than anticipated, but seemingly the best that could be achieved by considering only discourse entities and their relations.

The next stage of development focused on the extraction of short answers. The final result of this process is the set of heuristics described above for the individual question types. We proceeded to this task by categorizing the problems and the likely solutions.

In extending DIMAP-QA to extract 50-byte answers, we found that we could successfully identify appropriate phrases by greater attention to detailed syntactic and semantic structures within the sentence. We looked for opportunities for better characterization of syntactic and semantic roles played by constituents of the sentence; the appositive and genitive determiner constituents led to a significant improvement in performance, particularly for **who** and **what** questions. We were able to exploit this extraction with feedback to the sentence extraction (i.e., when a viable short answer was recognized, its sentence was given a higher score). This extension consisted of question-specific routines for extracting short answers based on the types of semantic relations in which the discourse entities participated. This improved scores to 0.740 for sentences and 0.493 for short answers.

At this point in development, it became clear that the model we were implementing could be characterized as just-in-time: improvements could be attained by implementing slight refinements taken from techniques like named-entity extraction and query expansion. It was only at this point in development that the lexical resources were integrated. Although these resources could have been used directly to answer questions, the just-in-time model used them instead for substantiation. For example, in "where" questions, definitions provided a background set of discourse entities used in evaluating document sentences. For "what" questions (e.g., "what country"), dictionary definitions were examined to determine whether a document discourse entity was defined as or had the hypernym "country". If no match, the thesaurus was examined to determine if the hypernym for a document discourse entity was in the same thesaurus category (e.g., as "country" where "Belgium" is defined as a "kingdom").

The final set of improvements to DIMAP-QA came from a more detailed evaluation of the short answers. These changes can be characterized as reflecting a more global view of the questions, identifying their critical components and implementing procedures for decreasing the scores of sentences that were given inappropriately high scores.

Incorporation of the lexical resources and the further evaluation of the short and sentence answers in light of the key words in the questions improved the TREC-8 scores to 0.803 for sentences and 0.597 for short answers. It was at this point that the system was frozen for the participation in TREC-9.

In examining the TREC-9 results, we have taken a similar approach to categorizing the failures. In general, we have found that there is nothing qualitatively different from our performance with the TREC-8 questions. We have, for the most part, extracted appropriate sentences for detailed analysis

(96.5%). Availability of the appropriate document is the most prevalent problem (34% of the failures). About 12% of the failures can be attributed to the need to degrade the scores of too highly scored sentences. Another 10% require improved characterization and extraction of constituents from the parse output. About 10% of the questions can be answered by improved routines for interacting with the lexical resources. About 6% can be characterized as difficult problems. The remaining problems seem to require better examination of the question components or modification of the algorithms for the individual questions. The routines that were implemented for the specific question types need to be evaluated for how well they work together (i.e., as some routines were implemented, they may have degraded other routines).

As mentioned earlier, we experienced significant problems with processing *Associated Press*, *Wall Street Journal*, and *San Jose Mercury News* documents. We reran the entire 10- and 20-document sets after our formal submission and estimate that these problems reduced our overall performance by about 0.028.

In Table 4, the adjusted score for the 20-document run was 0.394, compared to 0.412 for the 10-document run. This indicates that we actually experienced a degradation in performance in going from 10 to 20 documents. Overall, in examining the official scores, looking for cases where we performed better on the 10 document set than the 20 document set, we found that this amounted to 0.042 loss of points.

6. Anticipated Improvements

The immediate possibilities for improvements are many and the possibilities for exploration are quite diverse. In addition, there are opportunities to be explored for integrating DIMAP-QA within more generalized search engines.

The clearest avenue of improvement is indicated by the question variations in questions 701 to 893. For 16 variants we were unable to answer in the base 500 questions because the appropriate documents were not in the top 10; the problem persisted for 8 questions. Of the remainder, we were able to obtain an answer under 2 variations. Of the other 38 variation sets, we did not obtain answers for 18 of the base questions, but were able to find answers in one or more of the variations for 11 sets. This suggests that improvements may be obtained by finding the "best" canonical form for a question. (For most of the variants, the reformulated questions gave rise to quite different document positions of appropriate documents, underscoring again the significance of the retrieval problem.)

The use of the dictionary and thesaurus in this year's system was quite rudimentary. Analyzing the questions, we found that 35% were either definitional, answerable by dictionary lookup, or supportable by the dictionary. Implementing procedures similar to those used in answering **where** questions will lead to substantial improvements.

7. Summary

The CL Research system was reasonably successful in answering questions by selecting sentences from the documents in which the answers occur. The system generally indicates the viability of using relational triples (i.e., structural information in a sentence, consisting of discourse entities, semantic relations, and the governing words to which the entities are bound in the sentence) for question-answering. Post-hoc analysis of the results suggests several further improvements and the potential for investigating other avenues that make use of semantic networks and computational lexicology.

Acknowledgements

We want to thank Richard Tardif of Macquarie for making Big Mac and the Macquarie thesaurus available, Steve Green for his perl scripts for converting the Macquarie thesaurus into WordNet format, and Thomas Pötter for directing us to MetaKit and making us think hard about our approach.

References

Fellbaum, C. (1998). *WordNet: An electronic lexical database.* Cambridge, Massachusetts: MIT Press.

Green, S. J. (Stephen.Green@East.Sun.COM). (2000, 27 January). (Macquarie thesaurus).

Litkowski, K. C. (2000). Question-Answering Using Semantic Relation Triples. In: Voorhees, E. M. & Harman, D. K. (eds) *The Eighth Text Retrieval Conference (TREC-8)*, NIST Special Publication 500-246, 349-356.

Richardson, S. D. (1997). Determining similarity and inferring relations in a lexical knowledge base [Diss], New York, NY: The City University of New York.