CL Research Experiments in TREC-10 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-10 only slightly extends its semantic relation triple (logical form) technology in which documents are fully parsed and databases built around discourse entities. Time constraints did not allow us to make various changes planned from TREC-9. TREC-10 changes made fuller use of the integrated machine-readable lexical resources and extended the question-answering capability to handle list and context questions. Experiments to further exploit the dictionary resources were not fully completed at the time of the TREC-10 submission, affecting planned revisions in other QA components.

The official score for the main TREC-10 QA task was 0.120 (compared to 0.135 in TREC-9), based on processing 10 of the top 50 documents provided by NIST, compared to the average of 0.235 for 67 submissions. Post-hoc analysis suggests a more accurate assessment of DIMAP-QA's performance in identifying answers is 0.217. For the list task, the CL Research average accuracy was 0.13 and 0.12 for two runs compared to the average of 0.222. For the context questions, CL Research had mean reciprocal rank score of 0.178, 5th of the 7 submissions.

1. Introduction

TREC-10 DIMAP-QA proceeded from last year's version (Litkowski, 2001) primarily by attempting to integrate dictionary definition lookup into **what** and **who** questions, extending our success from using definitions in handling **where** questions using the dictionary. However, our strategy for **where** questions did not generalize, in part because of the poor retrieval performance of the NIST top documents when dealing with definition questions. We also added mechanisms for answering list questions, involving only a test for a numerical term in the question, but keeping the remaining functionality the same as for **what** questions, just returning the number of answers required. For context questions, we made no changes whatever to the system, yet still managed to obtain results consistent with our general question answering, even for the later questions of a given set.

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 QA track provides an opportunity for experimenting with question answering using syntactical clues and semantic evidence from use of computational lexical resources (dictionary and thesaurus).

2. Problem Description

Participants in the main TREC-10 QA track were provided with 500 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 in the list task were given 25 questions, each of which required a specified number of answers; the top 50 documents were also provided. Participants in the context task were given 10 question sets, varying in number from 3 to 9 questions; the top 50 documents retrieved using the first question of each set were also provided.

---

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

Participants in the main were required to answer the 500 questions in 50-byte answers. For each question, participants were to provide 5 answers, with a score attached to each for use in evaluating ties.[2] In TREC-10, a valid answer could be NIL, indicating that there was no answer in the document set; NIST included 49 questions for which no answer exists in the collection. For the list questions, participants were to return exactly the number of answers specified in the question. For the context questions, 5 answers were to be provided for each question of the set; the questions were constructed in a way so that later questions of the set depended on the answers to the earlier questions. NIST evaluators then judged whether each answer contained a correct answer. Scores were assigned as the inverse rank for the main and the context tasks. 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. If a NIL answer was returned, and was deemed valid, its position in the ranked list of answers was used as the rank. The final score was then computed as the average score over the entire set of questions. For the list questions, the "average accuracy" was computed as the number of correct answers divided by the number of required answers.

CL Research submitted 5 runs, 2 each for the main task and the list task and one for the context task. For the main and list tasks, one run analyzed 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. For the context task, only the top 10 documents were included in attempting to answer each of the questions in the set.

3. System Description

The CL Research question-answering 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 answers for submission. For the list task, the specified number of answers was submitted.

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.

For TREC-10, the top 20 documents (as ranked by the search engine) were analyzed for the main task, with one database containing only the processing for the top 10 documents and the other for the full 20 documents. Overall, this resulted in processing 9889 documents from which 225,248 sentences were identified and presented to the parser. Thus, we used an average of 22.8 sentences per document (down from 28.9 in TREC-9 and 31.9 in TREC-8) or 228 sentences for the 10-document set and 456 for the 20-document set.

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

[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.

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-10, parsing exceptions occurred for only 543 sentences out of 225069 (0.0024, up from 0.0002), with another 179 "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 10,916 (4.8 percent, down from 7.9 percent). For TREC-10, we were able to make use of the integrated dictionary to dynamically create entries for the parsing dictionary.

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. This involves 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).

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-10, 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,2,174,332 semantic relation triples were created in parsing the 225,248 sentences, an average of 9.7 triples per sentence (about the same as in TREC-9).

The same functionality was used to create database records for the 500 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 1576 triples, an average of 3.15 triples per question. This is down from 3.3 per question in TREC-9 and 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 component of the question-answering system is an integrated 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 (The Macquarie Dictionary 1997), 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). For TREC-10, the entire dictionary was reparsed to reflect improvements in the semantic creation techniques since TREC-9.

The Macquarie thesaurus is 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. With a set of perl scripts, the thesaurus data has been inverted 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-10, a database of documents was created for each question, as provided by the NIST generic search engine. A single database was created for each question in the main task, the list task, and one overall database to handle each of the questions in each context set. 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) detailed analysis of the question to set the stage for detailed analysis of the sentences according to the type of question, (2) coarse filtering of the records in the database to select potential sentences, (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. Few changes were made in each of these steps from TREC-9, so the description is largely the same, with some discussion of changes planned but not implemented in time for TREC-10.

3.5.1 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. **What** questions were further analyzed to determine if they have a number modifying the head noun, in which case these were treated as list questions. (A few questions in the main task were thereby turned into list questions, limiting the number of answers returned.)

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 **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 both **who** and **what** definition questions, an attempt was made to find the key noun in the Macquarie dictionary, creating a list of content words in its definitions for comparison with discourse entities in the sentences. 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 **size** questions, we identified the "size" word (e.g., "far" in "how far"). For **number** questions, we also looked for a year restriction.

3.5.2 Coarse Filtering of Sentences

The second step in the question-answering phase was the development of an initial set of sentences. In previous years, this was the first step, but with the addition of definition lookup as part of the analysis of question type, this was moved. Basically, the discourse entities in the question records are used to filter the records in the document database. However, this list is extended when a "definition" question is recognized, by adding words from the definition as obtained from the dictionary.[3] 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. In TREC-10, each hit of a discourse entity in a sentence added a score of 5 points to the sentence; this score determined the order in which sentences were further evaluated. Sentences with MWU discourse entities having a question discourse entity as a substring were selected during this screening, but were given no points and hence examined last in the detailed evaluation of the sentences.

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

[3] In TREC-9, there were 35 "definition" questions. In TREC-10, the number increased to 165 (including **where** questions). In addition, 43 questions were identified as susceptible of "dictionary support", where the answer could be looked up in the dictionary, with the expectation that the question elements would be discernible in the definition of the answer.

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. For **who**, **what**, and **location** definition questions, a background array of content words from the definitions was developed for later comparison with the answer.

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.

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 measu2ement 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.

Also, if either of these two question types was a definition question, we added points for each discourse entity that was among the content words of the definition.

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, with the 50-byte answer extracted from the original sentences retrieved from the documents.

4. TREC-10 Q&A Results

CL Research submitted 2 runs for the main task; the official scores for these runs are shown in Table 1. The score is the mean reciprocal rank of the best answer over all 492 questions that were included in the final judgments. The score of 0.120 for run clr01b1 means that, over all questions, the CL Research system provided a sentence with a correct answer at the 8th position. This compares to an average score of 0.235 among all submissions for the TREC-9 QA 250-byte answers (i.e., a correct answer slightly worse than the 4th position).

Table 1. CL Research Run Scores

Run	Doc. Num.	Type	Score	TREC Ave.
clr01b1	10	50-byte	0.120	0.235
clr01b2	20	50-byte	0.114	0.235

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 492 questions.

Table 2. Answers Found (492)

Run	Doc. Num.	Type	Num	Pct.
clr00b1	10	50-byte	94	0.191
clr00b2	20	50-byte	96	0.195

For the list task, the CL Research average accuracy was 0.13 and 0.12 for two runs compared to the average of 0.222. For the context questions, CL Research had mean reciprocal rank score of 0.178, 5th of the 7 submissions.

5. Analysis

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 3. Highest ranked top document containing strict answer string

Document Number	Number of Questions
1-10	311
11-20	26
21-30	13
31-40	15
41-50	5
None	122

Of the 122 questions having no answer in the top 50 documents, 49 have been judged as having no answers in the document collection. Adjusting the scores to include only questions that might have been answered (311 in the 10 document analysis and 337 in the 20 document analysis), the CL Research performance, shown in Table 4, is somewhat increased. For the 10-document case, the result is 0.217, compared to the average score of 0.235, while for the 20-document case, the adjusted result is down to 0.176.

Table 4. CL Research Adjusted Run Scores				
Run	Doc. Num.	Type	Score	TREC Ave.
clr01b1	10	50-byte	0.217	0.235
clr01b2	20	50-byte	0.176	0.235

A significant malfunction occurred from a program bug affecting the 20-document runs, where only two answers were submitted for the majority of questions. Notwithstanding, our system performed less well when additional documents were analyzed. 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 and 3.15 in TREC-10. 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. In particular, with the large number of definition questions (estimated at 165 of 492), retrieval based solely on the word to be defined is much less likely to obtain a document with the definition.

We examined our results using 250-byte answers as well. For the 10-document case, we obtained a score of 0.296 unadjusted and 0.465 adjusted. The difference in results indicates that we are generally narrowing down the candidate sentences, but having difficulty picking out the answer string.

For TREC-9, CL Research experimented with the Macquarie dictionary in support of answers to **location** questions. This strategy worked reasonably well in TREC-10, where we obtained an adjusted score of 0.319 for this type of question. However, it did not work for **what** and **who** definition questions. Part of this failure can be attributed to our mechanism for ranking, where we had not yet implemented an adequate test for the correctness of an answer. We have made some initial changes in our strategy that clearly lead to an improvement, but we have not yet been able to assess the overall effect of these changes.

We have not yet been able to complete our characterization of failures for TREC-10. In general, the problems lie in not being able to eliminate sentences that have a lot of hits with the discourse entities in the questions, giving too much weight to this aspect. The effect is that as we add further documents, sentences not containing the correct answer are given undue weight, crowding out sentences that contain the answer. In addition, our strategy for evaluating phrases within a sentence suffers from the same difficulty, giving too much weight to the wrong discourse entities.

6. Anticipated Improvements

As indicated earlier, we are in the process of making many changes to our question-answering system and these were not completed in time for our submission.

We are in the process of extending our document processing to incorporate discourse analysis techniques, building on the discourse entities. These changes will characterize the discourse entities semantically, in addition to resolving anaphor and definite references. Discourse structure (the relation of segments to one another) will also be captured. This amounts to tagging a document with semantic classes, named-entity types, and discourse relations over sentence spans longer than noun phrases. A key component in these characterizations will be the integrated use of WordNet and the Macquarie dictionary and thesaurus.

At the same time, we have been modifying our question-answering strategies to home in on semantic types and syntactic structures more likely to provide the answers. Initial results with definition questions show considerable improvement over our TREC-10 results. The discourse analysis has proved useful in making modifications to these QA strategies. The reverse has also proved to be the case, namely, that the QA strategies inform the manner in which we perform the discourse analysis.

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.

References

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

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. Gaithersburg, MD., 349-356.

Litkowski, K. C. (2001). Syntactic Clues and Lexical Resources in Question-Answering. In E. M. Voorhees & D. K. Harman (eds.), *The Ninth Text Retrieval Conference (TREC-9)*. NIST Special Publication 500-249. Gaithersburg, MD., 157-166.

The Macquarie Dictionary (A. Delbridge, Ed.). (1997). Australia: The Macquarie Library Pty Ltd

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