Syntactic Clues and Lexical Resources in Question-Answering

Ken Litkowski
CL Research
ken@clres.com
http://www.clres.com
http://www.clres.com/trec9/index.html

CL Research QA System

- Sentence splitter that separated the source documents into individual sentences
- Parser which parsed each sentence into a parse tree containing the constituents of the sentence
- Parse tree analyzer to identify important elements and create semantic relation triples stored in a database
- Question-answering program
  - parsed the question into semantic relation triples, except with an unbound variable
  - matched the question database records with the document database to answer the question

Semantic Relation Triples

Full document parsing identifies discourse entities and their semantic roles (logical forms, conceptual graphs)

- **Discourse entities**
  - Noun constituents, numbers, adjective sequences, possessives, leading noun sequences, ordinals, time phrases, predicative adjective phrases (no need for semantic relation or governing word)
- **Semantic relations**
  - Characterizes entities' semantic roles in sentence
  - Agent, theme, location, manner, modifier, purpose, time, relative clauses, appositives (SUBJ, OBJ, prepositions heading prepositional phrases)
- **Governing words**
  - The word in the sentence that the discourse entity stood in relation to
  - For "SUBJ," "OBJ," and "TIME," the main verb of the sentence
  - For prepositions, the noun or verb that the prepositional phrase modified
  - For the adjectives, numbers, and clauses, the noun that was modified

TREC-9 Processing Summary

- Only the top 20 documents (as ranked by the NIST search engine) were parsed (up from 10 in TREC-8)
- Processed 14,605 (from 1977) documents from which 422,562 (63,118) sentences were identified and presented to the parser
- An average of 28.9 sentences per document or 580 sentences in attempting to answer each question
- 4,149,106 (467,889) semantic relation triples were created in parsing the sentences, an average of 9.8 (7.4) triples per sentence
- 2272 triples (for 693 questions), an average of 3.3 triples per question (down from 4.5 last year)
- "Simpler" questions, more difficult to answer (less information to match)

Lexical Resources

Macquarie Dictionary and Thesaurus

- Dictionary
  - Provided in machine-readable form, converted into DIMAP format, and uploaded into DIMAP dictionaries (120,000 headwords), capturing all dictionary data and creating machine-tractable additions (typical subject, regular expression phrases, etc.)
  - 270,000 definitions parsed to populate dictionary with semantic relations links (hypernym, synonym, typical subjects and objects, semantic components such as manner, purpose, class membership and inclusion (similar to MindNet)
- Thesaurus
  - Roget-style 812 categories broken into paragraphs (>1 for each part of speech) and subparagraphs
  - Inverted into alphabetical order by part of speech with links to subparagraph level (facilitating lookup, similar to WordNet)

Document and Question Database Development

- Analyzed the parse tree of each sentence and question to
  - Extract numbers, adjective sequences, possessives, leading noun sequences, ordinals, time phrases, predicative adjective phrases, conjuncts, and noun constituents as discourse entities, and multiword units
  - Capture the semantic roles of the entities, as generally understood in linguistics, including agent, theme, location, manner, modifier, purpose, and time
  - Capture syntactic constituents (relative clauses, apposivitives, parentheticals, pre-appositives, genitive determiners, gerundial post-modifiers)
  - Identify the governing word for the discourse entity (the main verb, attachment point for prepositions, or noun modified)
- Some question rephrasing
- Improved parsing robustness (0.0002 errors, down from 0.0080)
- Improved semantic relation extraction (32 percent increase per sentence, 9.8 from 7.4 per sentence)

Question Answering Routines

Matching the database records for an individual question against the database of documents for that question

- **Coarse filtering of the records in the database to select potential sentences**
- **Identification of key question elements (key noun, verb, and adjective modifier), year restrictions, more refined filtering of the sentences according to the type of question**
- **Extraction of short answers (question-specific routines)**
- **Evaluation of sentence and short answer quality**
  - Good short answer increasing score of sentence
  - Looking for derived forms (”assassination” from “assassinate”)
  - Number of hits in appositives (”who” and “what” questions)
  - Favoring proper noun answers (including comparison of definition proper nouns with context)
  - Looking for measurement (”unit”) words for “size” questions

Extraction of Short Answers (1)

- **Time questions** (“when”, “what was the year” or “what was the date”): presence of TIME semantic relation (automatically assigned by parser, such as “last Thursday” or “in 1972"), with discourse entity containing an integer or having a word marked in the parser's dictionary as representing a time period, measurement time, month, or weekday
- **Location questions** (”where”): presence of “in”, “at”, “on” prepositional phrases, capitalized word modifying key noun, genitive determiners, context containing capitalized words from key noun’s dictionary definition
- **Who questions** (“who” or “whose”): search for copular relations or testing non-copular verbs for nouns in relation indicated by unbound variable, looking for appositive hits

Extraction of Short Answers (2)

- **What questions** (”what” or “which,” used alone or as question determiners, and unclassified questions, such as “why” or “name the”): search for copular relations or testing non-copular verbs for nouns in relation indicated by unbound variable, looking for appositive hits, testing discourse entities against dictionary definitions, hypernym matches, or common thesaurus category
- **Size questions** (“how” followed by an adjective): record that has a NUM semantic relation, examining governing word for measure, a unit, or a measurement size, with discourse entity and governing word as answer
- **Number questions** (“how many”): record that has a NUM semantic relation with the discourse entity in the question as the governing word

Use of Lexical Resources

- Analysis of multiword question discourse entities against dictionary to determine named-entity and common noun phrases
- Background of capitalized words in key noun's definition for comparision with context in "where" questions
- Looking for derived forms of verbs (nominalizations such as "assassination" from "assassinate")
- For "what" questions, comparing definitions of discourse entities for key noun, hypernym matches, and presence in same thesaurus category (e.g., "Belgium" for "what country" where "country" is not in definition, but rather "kingdom" is its hypernym, with "country" and "kingdom" in same thesaurus category)
- Looking for "unit" in a potential measurement word for "size" questions

TREC-9 QA Results

Official Results

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

Adjusted Scores for Documents Attempted

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

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

- **Rapidly decreasing returns**
- **Simple questions have less information**
- **Suggest use of dictionary lookup to analyze questions and feedback to document retrieval**

Post-Hoc Analysis (1)
(Changes from TREC-8)

- Dealing with problems from last year
  - Unprocessed documents, resolving parsing problems in documents and questions, and resolving triple extraction problems moved sentence answers from last year's 0.281 to 0.550
- Extaction of short answers
  - Implementation of question-specific routines for extracting short answers (including just-in-time anaphora resolution), with feedback to evaluation of sentence answers (i.e., increasing scores) improved TREC-8 scores to 0.740 for sentences and 0.493 for short answers
- Integration of lexical resources
  - Selective use of integrated dictionary and thesaurus (testing of capitalization and examination of definitions, hypernyms, and thesaurus categories) improved TREC-8 scores to 0.803 for sentences and 0.597 for short answers

Post-Hoc Analysis (2)
(Failure Identification)

- **Document inclusion**: affects 34% of questions, requires more effective document retrieval (use of dictionary in posing question to retrieval engine)
- **Ranking of answers**: only 3.5% failure in extracting appropriate sentence, so remaining failures due to ranking and extraction mechanisms
  - Degrading scores of high-ranked answers (12%)
  - Improved characterization and extraction of constituents from parse output (10%)
- **Sentence splitting**: unfamiliarity with nuances of markup reduced performance by 0.028
  - Effect on adjusted score would be increase to 0.440

Post-Hoc Analysis (3)

(Analysis of Other Wrongly-Answered Questions)

- Question variations (701 to 893)
  - 16 question variation sets in base had no answer in top documents (8 in variations, with 2 now answered)
  - 38 other question variation sets (18 unanswered in base, 11 answered in variations)
  - Suggestion of canonical form for questions
  - Reformulated questions gave rise to quite different "top documents", underscoring significance of retrieval problem
- Considerable improvement still likely from improved syntactic characterization of parse output and extraction of triples
- Likely improvements from use of definitions in analyzing "who" and "what" questions
- Small percent of failures due to anaphora and cataphora

Conclusions and Future Work

- **Semantic relation extraction still a viable method**
- **Retrieval vs. Question-Answering**
  - **"Simple" questions make retrieval a problem**
  - **Need for a feedback mechanism**
  - **Dictionary lookup in question analysis can enhance retrieval**
- **Improvement of extraction techniques, particularly enhancing and generalizing semantic relations**
- **Further exploitation of lexical resources**
  - **Targeting now possible to optimize their integration**
  - **Use for question-answering (and other applications) can guide construction of appropriate lexical data**