TREC 2005
Systems Overview

In the TREC2005 QA track, there were 50+ entries (depending on source). There were two separate tasks: the main task and the relationship task. In the main task, questions were grouped by target (topic), additional topic type in 2005, events. Questions types: factoid (requiring a single short answer and its supporting reference), list (a group of factoids of the same type, in essence), and other (definitions, any other information found). Here is a summary of the accuracy scores for the main task.

<table>
<thead>
<tr>
<th>Question type</th>
<th>Factoid</th>
<th>List</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top score</td>
<td>0.713</td>
<td>0.468</td>
<td>0.248</td>
</tr>
<tr>
<td>Median scores</td>
<td>0.152</td>
<td>0.053</td>
<td>0.156</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.014</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Ten representative systems are described below. The approaches vary widely.

**QACTIS**
Primary approach: attributed entity-relations graphs
Question processing: append topic for “anaphora resolution”, determine response type
Document retrieval: Lemur, with redundancy elimination (same Lemur score, same content, eliminate one)
Passage Selection: WordNet for hypo/hypernym relationships (limited value with named entities), Semantic Forests (proper noun “dictionaries”, 300 categories, not big enough), Wikipedia to augment (off-line access, captured “snapshots”, small gain ~1%), issue finding reference in AQUAINT corpus
Accuracy: 0.257, 0.103, 0.241

**U Sheffield**
Primary approach: multiple approaches
Question processing: SUPPLE, specific question grammars \(\Rightarrow\) semantic representation of question, hand-crafted table lookup to get expected answer type. Create independent questions, pronominal resolution, anaphora issues (“the first flight” / “space shuttles” — “the center” / “Berkman Center for Internet and Society”)
Document retrieval: 1) Lucene, 2) MadCow (in-house Boolean search engine)
Lucene: documents separated into annotated paragraphs and indexed
MadCow: semantic filtering based on answer type,
Passage selection: 1) shallow processing based on semantic typing, 2) syntactic analysis and logical form matching
Semantic entity detection and data (date & number) normalization
WordNet for question expansion: problematic, overgeneration, esp with NN
assumption, each answer is contained in single sentences, no coreference resolution, eliminate if 70% overlap with extracted sentences (shallow methods only for redundancy detection)

Accuracy: 0.110 (0.202), 0.035, 0.158

**Jellyfish**

Primary approach: regular expression rewriting

Question processing: complete questions using substitution with `<TARGET>`, add metadata for question category (answer type, units)

Document retrieval: 1) PRISE search engine, 2) MySQL full-text search

Passage selection: regex to mark `<TARGET>`, extract these sentences, no handling of intra-sentence references

regex to mark potential answers based on question type (date, country), uses predefined lists of potential values.

Answer retrieval: regex to match annotated passages and question metadata, no ranking step.

Accuracy: 0.110, 0.033, 0.088

**DLT**

Primary approach:

Question processing: POS tagging using Xelda, recognize question constructs (11 of 82 prominent), append target for anaphor resolution, weight identified constructs

Passage selection: create Boolean queries, AND them together, (enhance using Local Context Analysis?), Lucene index individual sentences, get n best (30), if not n results, relax query by removing least significant term.

Answer retrieval: NER on results, using grammar / exhaustive lists (in-house), score using co-occurrence of key phrases, question weights, distance from NE.

Accuracy: 0.177

**TALP-UPC**

Primary approach: voting between last year’s, this year’s, and Web

Last Year’s


Passage selection: Lucene, indexed whole AQUAINT corpus and idf weights, also POS tagging, lemmas and NER (both lemmatized and original) on whole corpus.

Segmentation in to individual sentences, scoring on semantic content: \( tf \times idf \). Filter based on mandatory constraints, if below threshold, relax selection cutoff, retry.
Answer retrieval: Support Vector Machine (SVM) trained on TREC8 – TREC12 QA corpora and published answers, scoring based on relaxation level that allowed the extraction, rule score for extraction, semantic score, passage score.

This Year’s

Question Processing: expected answer type, keywords, similar to last year, additionally unigrams and bigrams for question words, unigrams and bigrams for phrase heads, n-grams expanded by thesaurus. Max Ent to classify results.

Passage retrieval: similar to last year

Answer extraction: heuristic based: same word sequence, punctuation flag (clause terminal), comma words for appositives, same sentence, matched keywords, distance.

Web-based

Identify common question patterns, knowledge mining based on the assumption the answer and question will share similar term structure

Accuracy: 0.172

**Language Computer Corp**

Primary approach: syntactic parsing, NER, reference resolution

Question processing: extensive co-reference resolution. Determine answer type, select keywords for passage retrieval.

  Target 136: Shiite
  Q136.1: Who was the first Imam<1> of the Shiite sect of Islam?
  Q136.2: Where is his<1> tomb?
  Q136.3: What was this person’s<1> relationship to the Prophet Mohammad?
  Q136.4: Who was the third Imam<2> of Shiite Muslims?
  Q136.5: When did he<2> die?

Passage selection: scant details provided, includes ranking of passages.

Answer processing: scant details, accuracy boosted using the Web (redundancy boosts Web answers), Cogex logical prover to resolve ambiguities, rerank answers (semantic analysis)

Accuracy: 0.713, 0.468, 0.228

**QED**

Primary approach: Combinatorial Categorial (Categorical?) Grammar

Question processing: CCG analysis to get Discourse Representation Structure (DRS)

Document retrieval: preprocess AQUANT. Lemur to extract documents based on target. Broken into two sentence pages. Select if at least one word of target phrase.

Passage selection: POS tagging, NER, CCG analysis of retrieved documents,

Answer retrieval: DRS unification (relaxed), reranking based on Google API to Web resources.
Independent off-line processing; restructure whole corpus around potential answer types, so far, *person*, *location*, *organization*. Augment QED answers from this cache. NB: decrease in accuracy using this system.

Accuracy: 0.215

**U Amsterdam**

Primary approach: multi-stream extraction, XML representation

Preprocess: Collect “hard” question data (birthplaces of people, groups and membership, nicknames, organizations and their founders/founding dates). Store in tables. Break AQUAINT corpus into paragraph sequences. Annotate with token boundary, syntactic, NE info. Store in XML.

Question processing: parse using Charniak to extend questions, resolve NP/PP chunks.
NERC: addressed problems with NER using post processing, personal names within organizational names and titles misclassified as organizations.

Document retrieval: annotation matching

Answer retrieval: scoring centroid based

Accuracy: 0.201

**ILQUA**

Primary approach: NE-tagged passages

Question processing: question type, NE type, WordNet supplement with morphological forms and verbal synonyms. No noun synonyms, too much noise.

Document retrieval: Inquery (Amhearst), filter on answer type, question terms, topic terms. Clustering didn’t help.

Passage selection: surface pattern matching. Based on POS tagging and question word: *when_be_NP_VP*

Answer retrieval: merge similar patterns and rank

Accuracy: 0.273, 0.12, 0.206