Question Answering by Diggery at TREC-2006

Stan Tomlinson, Ph.D.
stan@tomlinson.com

Diggery is a research and software project, investigating the extraction of concepts from well-written documents, with the idea of automating factoid search. The project is in its early to middle phases, and all information presented herein should be taken in light that this research is based on young software using new algorithms.

In January 2006, after significant tuning, the software could answer a few simple questions from small texts. Six months later, in July 2006, the first real exercise of the software on a non-trivially sized corpora was made for the TREC QA submission, and the software answered a few questions correctly. For this submission, only factoid questions were attempted.

Overview of Methods

The software, as implemented for TREC 2006, has three main components: indexing the words in the corpora, extracting concepts from source text for a particular topic, and answering questions from the extracted concepts.

A simple index engine was built that indexed, by sentence and document, the words in the corpora. For each question topic, the sentences which have the topic words were identified, and those sentences were passed to the concept extractor.

The concept extractor parses a source sentence and produces an intermediate syntactic tree. The constituent parts of the syntactic tree are compiled into concepts. If it is necessary to resolve anaphora, additional source sentences are parsed.

To answer a question, the question is parsed using the same underlying parser engine that is used to produce concepts. Using a computationally attractive algorithm, answers are derived.

The English parser is built with a hand-crafted non-deterministic context-free grammar and a custom compiler-compiler. The resulting parser supports statistical modeling. To handle anaphora in the QA questions, the parser's standard anaphora resolver was modified to utilize the topic words as a potential target.

The Run

For the TREC 2006 run, indexing was only rudimentarily optimized. It took about 5 days to index the entire corpora. For each topic, it typically would take between 2 and 7 minutes to look up in the index the documents that contain the topic, parse the sentences from the source documents, and perform the concept extraction. The query system would then be asked the questions. It would typically take under a minute to answer all the questions in a topic group. The system was reset, and the next topic was started.

The QA run, along with the research and software development, was performed on a mid-speed Pentium-class PC.

The Results

By my count, the software answered 9 out of 567 questions, 5 of them correctly (the judges count was both higher and lower, higher because some NIL answers were counted as correct, and lower because of non-specificity). Five correct is not an earth-shattering result. However, that it answered any questions correctly demonstrates the basic end-to-end framework is minimally functional, and provides a small proof-of-concept for the algorithms on a larger corpora.

Future Directions

Obviously, Diggery will need significant improvement before it even begins to approach alpha-level performance.

The vast majority of development time prior to submission was used creating building blocks, such as a dictionary, a parser, a grammar, the

extractor components, an inference engine, anaphora resolution, the answer algorithm components, and test platforms. These blocks were required before any substantial testing of the extraction and answering algorithms could be performed, and, although there remain large issues to be addressed, they are now stable. Some of the building-block issues are: the grammar used by the parser needs to be improved; the inference engine needs more rules; a database of common knowledge needs to be incorporated; anaphora resolution needs improvements; a noun-reference resolver will improve accuracy; an ellipsis resolver needs to be built; although dates, times, and verb tenses are accurately extracted from source sentences and queries, the information is not currently used when answering a question; speed increases are desirable.

Many of these issues can now be put on the back burner while the core research proceeds. Concept extraction and answer algorithm routines will be the primary focus of the research this next year.

Conclusions

The dearth of correct answers should not yet be perceived as an accurate indicator of future performance. As this short paper is being written, continued improvements are being made, and give hope that the algorithms will ultimately produce a robust system for accurately finding factoids and other information.