Watson: Trick Or Treat?

Drew McDermott
September 5, 2012

How close is IBM’s Watson program to a “general” intelligence? Or a general converser?
What Does Watson Do?

It plays the Jeopardy (insert exclamation point here) game.

This is a long-running U.S. game show. Mostly a trivia game, in which to answer the question you have to be the first to hit a buzzer.

So you have to judge how confident you are in your answer very quickly.
Question Answering

To answer a question, you have to retrieve text sources that contain the answer. (In a small percentage of cases, the answer comes from a structured knowledge source, like a list of U.S. presidents.)

The TREC conferences are (apparently) the major venue for testing algorithms in this country. (Foreign teams can enter.)
Recall/Precision Tradeoff

As in other information-retrieval (IR) situations, the more documents you find, the less likely to be useful they are.

Recall = \frac{\text{number of relevant docs found}}{\text{number of relevant docs in collection}}

Precision = \frac{\text{number of relevant docs found}}{\text{number of docs found}}
Confidence/Accuracy Tradeoff

- In Jeopardy, time to hitting the buzzer is crucial.

- Ignoring opponents, the issue is how much accuracy you sacrifice by answering when unconfident.
The Competition

Darker points are for Ken Jennings, superstar.
DEEPQA Baseline Performance

Figure 4. Baseline Performance.
What Is TREC?

TREC (Text REtrieval Conference) seems to be more of a competition. It’s been run annually by NIST (National Institutes of Standards and Technology) for over 12 years. See trec.nist.gov.

Competitors are given a domain and sample texts to work with. (There are often multiple tracks, looking at different domains.) Then they are given a set of documents and must answer questions based on information in those documents. Sometimes a set of questions is given, and every story that looks promising must be mined for answers to those questions (information extraction).
Reengineering QA

1. Many knowledge sources

2. Each must supply a confidence rating (of some kind). [How combined?]

3. Handling decomposition
Decomposition

Example:

“Of the four countries in the world that the United States does not have diplomatic relations with, the one that’s farthest north.”

You have to use the answer to part 1 (“What four countries does the US not have diplomatic relations with?”) to form the part-2 question: “Of the four countries Bhutan, Cuba, Iran, and North Korea, which is furthest north?”
Two Key Terms

**LAT**: Lexical Answer Type, the word in the clue that is the type of the answer. “Invented in the 1500s . . ., this [chess] maneuver involves two pieces . . .”: the LAT is “maneuver.”

**Focus**: The part of the question that, if replaced by the answer, yields a true declarative sentence. Often starts with “this” (e.g., “this maneuver”).
Re-architected System
What Do All Those Boxes Do?

1. *Question Analysis*: Finding “deep” and “shallow” linguistic structures, question “classification,” focus and LAT detection, relation detection, decomposition

2. *Hypothesis Generation*: Finding “answer-sized snippets” from search results, plugging back into question in place of focus to produce hypotheses. Primary search $\rightarrow$ candidate-answer generation.
What Do All Those Boxes Do? (cont.)


Boiling Things Down

1. *Output of primary search:* Roughly 250 documents with 85% chance of containing the answer.

2. *Output of candidate-answer generation:* “Several hundred candidate answers.”

3. *Output of soft filtering:* “Roughly 100 candidates”

4. *Output of hypothesis and evidence scoring:* One answer + confidence estimate (probability of correctness)
Discussion Question

What is the role of KR in Watson?
Ranking and Confidence Estimation

Used machine-learning algorithms on test problems with known answers.

I suppose confidence is based on how often Watson answer differs from known answer.

Different learners for different question classes (factoid retrieval, puzzles, puns, etc.)
Watson’s Blunder

On “final jeopardy” question, everyone must answer, even if they’re not confident about their answer.

On last question of Day 1 of its match against Jennings and Rutter, YouTube records how it did:

http://www.youtube.com/watch?v=Y2wQQ-xSE4s

The puzzle is, how come the right answer wasn’t generated? Or scored low? Is the grammar too hard?
Tricky Question

“Its largest airport is named for a World War II hero; its second largest, for a World War II battle.”
Concluding Question

Could the Watson team have thought a little more “generally” and come closer to artificial general intelligence?