Watson:
An Overview of the DeepQA Project

AI Magazine 2010

IMB TEAM:

Slides by Diana Inkpen
Jeopardy!

• U.S. TV quiz show that has been on the air since 1984.
• Rich natural language questions covering a broad range of general knowledge.
• It is widely recognized as an entertaining game requiring smart, knowledgeable, and quick players.
• All contestants must pass a 50-question qualifying test to be eligible to play.
• The first two rounds of a game use a grid organized into six columns, each with a category label, and five rows with increasing dollar values.
Examples

• **Category:** General Science
  • **Clue:** When hit by electrons, a phosphor gives off electromagnetic energy in this form.
  • **Answer:** Light (or Photons)

• **Category:** Lincoln Blogs
  • **Clue:** Secretary Chase just submitted this to me for the third time; guess what, pal. This time I’m accepting it.
  • **Answer:** his resignation

• **Category:** Head North
  • **Clue:** They’re the two states you could be reentering if you’re crossing Florida’s northern border.
  • **Answer:** Georgia and Alabama
Other questions

• More complexity, subclues
• Excludes audio-visual questions

• Watson
  – no ASR (Automatic Speech Recognition)
  – no live Internet access during game
Lexical Answer Type Frequency

![Graph showing the frequency of lexical answer types (LATs)]
Evaluation Metrics

• question-answering precision
• speed
• confidence estimation
• clue selection
• betting strategy

• Objective: money reward
Precision vs. Percentage Answered for Two Theoretical Systems:
Perfect confidence estimation (upper line) and no confidence estimation (lower line)
Champion Human Performance at Jeopardy
Baseline System

• Practical Intelligent Question Answering Technology (PIQUANT) (Prager, Chu-Carroll, and Czuba 2004), under development at IBM for 6 years.
• in top three to five Text Retrieval Conference (TREC) QA systems. Developed
• PIQUANT is a classic QA pipeline with state-of-the-art techniques aimed largely at the TREC QA evaluation (Voorhees and Dang 2005).
• PIQUANT performed in the 33 percent accuracy range in TREC evaluations.
• OpenEphyra, an open-source QA framework developed at CMU. On TREC 2002 data, OpenEphyra answered 45 percent of the questions correctly.
Text Search Versus Knowledge Base Search (Named Entities Recognition)
DeepQA

• massive parallelism
• many experts
• pervasive confidence estimation
• integration of shallow and deep knowledge
DeepQA Architecture
Roles of Modules

• Content Acquisition
• Question Analysis
  – Question Classification
  – Focus and LAT Detection
  – Relation Detection
• Hypothesis Generation
  – Primary Search
  – Candidate Answer Generation
• Soft Filtering
• Hypothesis and Evidence Scoring
• Answer Merging
Evidence Profiles for Two Candidate Answers

(Dimensions on x-axis and relative strength on y-axis)
Status

• After approximately 3 years of effort by a core algorithmic team composed of 20 researchers and software engineers with a range of backgrounds in:
  – natural language processing
  – information retrieval
  – machine learning
  – computational linguistics
  – knowledge representation and reasoning
Accuracy on Jeopardy! and TREC

- IBM’s 2005 TREC QA
- CMU’s 2007 TREC QA
- DeepQA performs on BOTH tasks 9/2008
- DeepQA prior to Adaptation
Scale and Speed

- Apache UIMA, a framework implementation of the Unstructured Information Management Architecture (Ferrucci and Lally 2004).
- Components in DeepQA as UIMA annotators: software components that analyze text and produce annotations or assertions about text.
- UIMA facilitated rapid component integration, testing, and evaluation.
- Early implementations of Watson: 1 processor, 2h to answer 1 question.
- Highly parallel: UIMA-AS, enables the scaleout of UIMA applications using asynchronous messaging.
- Scales Watson out over 2500 compute cores.
- UIMA-AS handles communication, messaging, and queue management necessary using the open JMS standard.
- The UIMA-AS deployment of Watson: run-time latencies of 3–5s.
- To preprocess the corpus and create fast runtime indices used Hadoop map-reduce framework.
Conclusions

Big and successful effort !!!

Big publicity for:

• **Computer Science**
  – **Artificial Intelligence**
  – **Natural Language Engineering**

• **Software Engineering**
Questions?

• Ask Watson!