IBM Watson: a real NLP Application

Ruihong Huang
Assistant Professor
Computer Science and Engineering
Texas A&M University
NLP Applications around us

- Email Spam Detection
- Apple Siri
- Google Translate
- Search Engines: google, yahoo
- ......
a Typical NLP Pipeline

Texts → Sentence Splitting → Tokenization → Parts-of-speech Tagging

→ Syntactic Parsing
→ Semantic Role Labeling

→ Named Entity Recognition

→ Coreference Resolution
→ Entity Linking

→ Relation Extraction
The dog saw a man in the park.
Other Tasks

- Paraphrasing
- Textual Entailment
- Word Sense Disambiguation
- Semantic Parsing
- ......
Applications

- Text Summarization
- Machine Translation
- Text Generation
- Event Extraction
- Question Answering
- Dialogue Generation
- ......
Language Technology

making good progress

mostly solved

Sentiment analysis
Best roast chicken in San Francisco!
The waiter ignored us for 20 minutes.

Coreference resolution
Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)
I need new batteries for my mouse.

Parsing
I can see Alcatraz from the window!

Machine translation (MT)
第13届上海国际电影节开幕...
The 13th Shanghai International Film Festival...

Information extraction (IE)
You're invited to our dinner party, Friday May 27 at 8:30

still really hard

Question answering (QA)
Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase
XYZ acquired ABC yesterday
ABC has been taken over by XYZ

Summarization
The Dow Jones is up
The S&P500 jumped
Economy is good

Dialog
Where is Citizen Kane playing in SF?
Castro Theatre at 7:30. Do you want a ticket?
Question Answering

- Closed-domain: reading comprehension
- Open-domain: web-based
Question Answering

- Understanding questions
- Retrieving and processing relevant texts
- Answer Generation
Quiz

- How big is the system?
- Is Watson Online?
- open-domain or closed-domain?
- What NLP techniques used?
Real-Time Game Configuration Used in Sparring and Exhibition Games

Analysis of natural language content equivalent to 1 Million Books
Watson – a Workload Optimized System

- 90 x IBM Power 750\(^1\) servers
- 2880 POWER7 cores
- POWER7 3.55 GHz chip
- 500 GB per sec on-chip bandwidth
- 10 Gb Ethernet network
- 15 Terabytes of memory
- 20 Terabytes of disk, clustered
- Can operate at 80 Teraflops
- Runs IBM DeepQA software
- Scales out with and searches vast amounts of unstructured information with UIMA & Hadoop open source components
- Linux provides a scalable, open platform, optimized to exploit POWER7 performance
- 10 racks include servers, networking, shared disk system, cluster controllers

\(^1\) Note that the Power 750 featuring POWER7 is a commercially available server that runs AIX, IBM i and Linux and has been in market since Feb 2010
The Jeopardy! Challenge: A compelling and notable way to drive and measure the technology of automatic Question Answering along 5 Key Dimensions

- **Broad/Open Domain**
- **Complex Language**
- **High Precision**
- **Accurate Confidence**
- **High Speed**

$200
If you're standing, it's the direction you should look to check out the wainscoting.

$1000
Of the 4 countries in the world that the U.S. does not have diplomatic relations with, the one that's farthest north

$600
In cell division, mitosis splits the nucleus & cytokinesis splits this liquid *cushioning* the nucleus.
Real Language is Real Hard

• Chess
  – A finite, mathematically well-defined search space
  – Limited number of moves and states
  – Grounded in **explicit, unambiguous** mathematical rules

• Human Language
  – Ambiguous, contextual and implicit
  – Grounded only in **human cognition**
  – Seemingly infinite number of ways to express the same meaning
Our Focus is on reusable NLP technology for analyzing vast volumes of as-is text. Structured sources (DBs and KBs) provide background knowledge for interpreting the text.
The Best Human Performance: Our Analysis Reveals the Winner’s Cloud

Each dot represents an actual historical human Jeopardy! game.

Top human players are remarkably good.

Computers?

Winning Human Performance

Grand Champion Human Performance

2007 QA Computer System

More Confident

Less Confident
Where did it acquire knowledge?

Three types of knowledge

Domain Data (articles, books, documents)

Training and test question sets w/answer keys

NLP Resources (vocabularies, taxonomies, ontologies)

Wikipedia
Time, Inc.
New York Time
Encarta
Oxford University
Internet Movie Database
IBM Dictionary
... J! Archive/YAGO/dbPedia...

Total Raw Content

Preprocessed Content

- 17 GB
- 2.0 GB
- 7.4 GB
- 0.3 GB
- 0.11 GB
- 0.1 GB
- 0.01 GB
- XXX

- 70 GB
- 500 GB
Watson’s Knowledge for Jeopardy!

Watson has analyzed and stored the equivalent of about 1 million books (e.g., encyclopedias, dictionaries, news articles, reference texts, plays, etc).

Watson also uses structured sources such as WordNet and DBpedia.
Automatic Learning by “Reading”

Volumes of Text → Syntactic Frames → Semantic Frames

Sentence Parsing → Generalization & Statistical Aggregation

Syntactic Frames:
- subject
- verb
- object

Semantic Frames:
- Inventors patent inventions (0.8)
- Officials Submit Resignations (0.7)
- People earn degrees at schools (0.9)
- Fluid is a liquid (0.6)
- Liquid is a fluid (0.5)
- Vessels Sink (0.7)
- People sink 8-balls (0.5) (in pool/0.8)

© 2009 IBM Corporation
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture

DeepQA generates and scores many hypotheses using an extensible collection of **Natural Language Processing, Machine Learning** and **Reasoning Algorithms**. These gather and weigh evidence over both unstructured and structured content to determine the answer with the best confidence.
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture

1. Initial Question Formulated: “The name of this monetary unit comes from the word for "round"; earlier coins were often oval”

2. Watson performs question analysis, determines what is being asked.

3. It decides whether the question needs to be subdivided.
Analyzing the question

Category: WORLD GEOGRAPHY

Clue: In 1897 Swiss climber Matthias Zurbriggen became the first to scale this Argentinean peak.

Step 1 Watson dissects the clue to understand what it is asking for.
Watson tokenizes and parses the clue to identify the relationships between important words and find the focus of the clue, i.e., this Argentinean peak.
Not *Just* for Fun

**Category:** Edible Rhyme Time

A long, tiresome speech delivered by a frothy pie topping

Diagram:
- Diatribe
- Harangue
- Meringue
- Whipped Cream

Answer: **Meringue**  **Harangue**

Some Questions require Decomposition and Synthesis

© 2009 IBM Corporation
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture

In creating the hypotheses it will use, Watson consults numerous sources for potential answers...

Watson then starts to generate hypotheses based on decomposition and initial analysis...as many hypothesis as may be relevant to the initial question...
Step 2  Watson searches its content for text passages that relate to the clue.

Using important terms from the clue, Watson performs a search over millions of documents to find relevant passages.

**Timeline of Climbing the Matterhorn**

* August 25: H.R.H. the Duke of the Abruzzi made the ascent with Mr. A. F. Mummery and Dr. Norman Collie, and one porter, Pollinger, junior. According to Mummery the weather was threatening, and, the Prince climbing very well, they went exceedingly fast, so that their time was probably the quickest possible. They left the bivouac at the foot of the snow ridge at 3.40 a.m., and reached the summit at 9.50. A few days afterwards the first descent of the ridge was accomplished by Miss Bristow, with the guide Matthias Zurbriggen, of Macugnaga.

The first known ascent of Aconcagua was during an expedition led by Edward Fitz Gerald in the summer of 1897. Swiss climber Matthias Zurbriggen reached the summit alone on January 14 via today's Normal Route. A few days later Nicholas Lanti and Stuart Vines made the second ascent. These were the highest ascents in the world at that time. It's possible that the mountain had previously been climbed by Pre-Columbian Incans.
Step 3 Watson analyzes the text passages and generates possible “candidate answers”.

Watson extracts important entities – so called “candidate answers” – from the documents. The focus is on coverage, which means that as much as possible is added (here, peaks, mountain ranges, people). At that stage, these are just possible answers to Watson.
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture

6. Watson then uses algorithms to “score” each potential answer and assign a confidence to that answer...

7. Watson uses Evidence Sources to validate its hypothesis and help score the potential answers.

8. If the question was decomposed, Watson brings together hypotheses from sub-parts.
In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

In May, Gary arrived in India after he celebrated his anniversary in Portugal.

Evidence suggests “Gary” is the answer BUT the system must learn that keyword matching may be weak relative to other types of evidence.
On 27th May 1498, Vasco da Gama landed in Kappad Beach.

In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

The evidence is still not 100% certain.
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture

1. Initial Question
2. Question & Topic Analysis
3. Answer Sources
   - Primary Search
   - Candidate Answer Generation
4. Question Decomposition
5. Hypothesis Generation
6. Hypothesis & Evidence Scoring
7. Synthesis
8. Merging & Ranking
9. Learned Models help combine and weigh the Evidence
10. Once Watson has ranked its answers, it then provides its answers as well as the confidence it has in each answer.
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture
IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE
IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE

Keywords: 1698, comet, paramour, pink, ...
AnswerType(comet discoverer)
Date(1698)
Took(discoverer, ship)
Called(ship, Paramour Pink)
...
IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE
IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE.

Keywords: 1698, comet, paramour, pink, ...

AnswerType(comet discoverer)
Date(1698)
Took(discoverer, ship)
Called(ship, Paramour Pink)

Primary Search Results

Candidate Answer Generation

Isaac Newton
Wilhelm Tempel
HMS Paramour
Christiaan Huygens
Halley’s Comet
Edmond Halley
Pink Panther
Peter Sellers

…
IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE.
IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE.
IN 1698, THIS COMET DISCOVERER TOOK A SHIP CALLED THE PARAMOUR PINK ON THE FIRST PURELY SCIENTIFIC SEA VOYAGE
Primary Searches are independent, and each Search Result is analyzed independently to generate candidate answers.
Evidence is Retrieved for each Candidate Answer independently

Evidence scoring is done independently

Evidence Retrieval

Evidence Scoring

[0.58 0 -1.3 ... 0.97]

[0.71 1 13.4 ... 0.72]

[0.12 0 2.0 ... 0.40]

[0.84 1 10.6 ... 0.21]

1) Edmond Halley (0.85)
2) Christiaan Huygens (0.20)
3) Peter Sellers (0.05)
Development System Timing – Before Scale Out

- Single-threaded computation
- Search indexes on disk
- Remote Sesame server

May 2009: Development System Times (average = 2300 seconds)
After first 8 months of Scaleout Work …

- Move everything into RAM
- Scale out components with UIMA-AS
- Distribute search

March 2009: Answer Times (average = 14.3 seconds)
12 more months of Scaleout Work …

- Pre-compute deep NLP analysis of entire text corpus
- Hammer on every computation outlier
- Expand cluster

March 2010: Answer Times (average = 2.6 seconds)

IBM Watson Playing in the Winners Cloud

Baseline 12/06

© 2011 IBM Corporation
The Core Technical Team*
Researchers and Engineers in NLP, ML, IR, KR&R and CL at IBM Labs and a growing number of universities

<table>
<thead>
<tr>
<th>Systems &amp; Speed</th>
<th>Core Algorithms</th>
<th>University Collaborations &amp; Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eric Brown</td>
<td>Radu Florian</td>
<td>Eric Nyberg (CMU)</td>
</tr>
<tr>
<td>Jerry Cwiklik</td>
<td>David Gondek</td>
<td>James Allen (UMASS)</td>
</tr>
<tr>
<td>Pablo Duboue</td>
<td>Aditya Klayanpur</td>
<td>Andy Schlaikjer (CMU)</td>
</tr>
<tr>
<td>Eddie Epstein</td>
<td>Hiroshi Kanayama</td>
<td>Nico Schlaefer (CMU)</td>
</tr>
<tr>
<td>Tong Fin</td>
<td>Adam Lally</td>
<td>Ed Hovy (USC)</td>
</tr>
<tr>
<td>Dan Gruhl</td>
<td>John Prager</td>
<td>Manas Pathak (CMU)</td>
</tr>
<tr>
<td>Bhavani Iyer</td>
<td>Tony Levas</td>
<td>Bruce Porter (UT)</td>
</tr>
<tr>
<td>Adam Lally</td>
<td>Chris Welty</td>
<td>Chang Wang (UMASS)</td>
</tr>
<tr>
<td>Burn Lewis</td>
<td>Lei Zhang</td>
<td>Pallika Kanani (UMASS)</td>
</tr>
<tr>
<td>Marshall Schor</td>
<td>Wlodek Zadrozny</td>
<td>Hideki Shima (CMU)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Barbara Cutler (RPI)</td>
</tr>
<tr>
<td></td>
<td>Dafna Sheinwald</td>
<td>Boris Katz (MIT)</td>
</tr>
<tr>
<td></td>
<td>Siddharth Patwardhan</td>
<td>Alessandro Moschitti (Trento)</td>
</tr>
</tbody>
</table>

*NOT full-time Equivalents. Names listed if contributed some time to that part of project.

There is a broader team that contributed to delivering Watson for the “Stage”, to compete in Jeopardy Games.

Strategy
- David Gondek
- Jon Lenchner
- Gerry Tesauro
- James Fan
- John Prager

Speech
- Andy Aaron
- Raul Fernandez
- Miroslav Novak
- Andrew Rosenberg
- Roberto Sicconi

Data Annotation
- Karen Ingraffea
- Matt Mulholland

© 2011 IBM Corporation
Watson-enabled patient-centered healthcare solutions

Longitudinal Patient Electronic Health Information

- Patient Inquiry
- Patient Workup
- Differential Diagnosis
- Specialty Diagnosis & Treatment Options
- Caregiver Education
- Coding Automation
- Treatment Options
- Treatment Authorization
- Treatment Protocol Analysis
- Care Consideration Analysis
- What's New?
- Second Opinion
- On-going Treatment
- Specialty Research
- Population Analysis & Care Mgmt
- Genomic-based Analysis

Patient Lay Caregiver...PA... Nurse Practitioner... Physician
Potential Business Applications

**Healthcare / Life Sciences**: Diagnostic Assistance, Evidenced-Based, Collaborative Medicine

**Tech Support**: Help-desk, Contact Centers

**Enterprise Knowledge Management and Business Intelligence**

**Government**: Improved Information Sharing and Security

© 2009 IBM Corporation