Natural Language for Communication

Watson Overview
Our Study Path Forward for “Natural Language for Communication”

• Groundwork:
  – Review of probability: Ch. 13
  – Probabilistic reasoning over time: Ch. 15.1-15.3
  – Language models: Ch. 22.1

• Natural language for communication: Ch. 23
IBM Watson

Slides from Watson Team, Presenter: Joel Farrell, IBM

IBM’s Watson…

**Why Jeopardy?**

The game of *Jeopardy!* makes great demands on its players – from the range of topical knowledge covered to the nuances in language employed in the clues. The question IBM had for itself was “is it possible to build a computer system that could process big data and come up with sensible answers in seconds—so well that it could compete with human opponents?”
IBM Watson’s project started 2007

- Project started in 2007, lead David Ferrucci
- Initial goal: create a system able to process natural language & extract knowledge faster than any other computer or human

- Jeopardy! was chosen because it’s a huge challenge for a computer to find the questions to such “human” answers under time pressure

- Watson was NOT online!

- Watson weighs the probability of his answer being right – doesn’t ring the buzzer if he’s not confident enough

- Which questions Watson got wrong almost as interesting as which he got right!
Watson – a Workload Optimized System

- 90 x IBM Power 750 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
This means Watson…

- Operates at 80 teraflops. The human brain is estimated to have a processing power of 100 teraflops (100 trillion operations per second).
- Has the equivalent in memory (RAM) that the Library of Congress adds in books and media over a 4 month period
- Can process 200 million times more instructions per second than the Space Shuttle’s computers.
- Parses within 3 seconds the equivalent of the number of books on a 700 yard long book shelf…and pick out the relevant information, and create an answer.
A Grand Challenge Opportunity

- Capture the imagination
  - The Next *Deep Blue*

- Engage the scientific community
  - Envision new ways for computers to impact society & science
  - Drive important and measurable scientific advances

- Be Relevant to Important Problems
  - Enable better, faster decision making over unstructured and structured content
  - Business Intelligence, Knowledge Discovery and Management, Government, Compliance, Publishing, Legal, Healthcare, Business Integrity, Customer Relationship Management, Web Self-Service, Product Support, etc.
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
What Computers Find Easier (and Hard)

\[
\ln((12,546,798 \times \pi)^2) / 34,567.46 = \ 0.00885
\]

Select *Payment* where *Owner* = “David Jones” and *Type(Product) = “Laptop”,

<table>
<thead>
<tr>
<th>Owner</th>
<th>Serial Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Jones</td>
<td>45322190-AK</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Serial Number</th>
<th>Type</th>
<th>Invoice #</th>
</tr>
</thead>
<tbody>
<tr>
<td>45322190-AK</td>
<td>LapTop</td>
<td>INV10895</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Invoice #</th>
<th>Vendor</th>
<th>Payment</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV10895</td>
<td>MyBuy</td>
<td>$104.56</td>
</tr>
</tbody>
</table>

David Jones ≠ Dave Jones
What Computers Find Hard

Computer programs are natively explicit, fast and exacting in their calculation over numbers and symbols....But Natural Language is implicit, highly contextual, ambiguous and often imprecise.

- Where was X born?

  One day, from among his city views of Ulm, Otto chose a water color to send to Albert Einstein as a remembrance of Einstein’s birthplace.

- X ran this?

  If leadership is an art then surely Jack Welch has proved himself a master painter during his tenure at GE.
Some Basic Jeopardy! Clues

- Category: ENDS IN "TH"
  - This fish was thought to be extinct millions of years ago until one was found off South Africa in 1938
  - Answer: coelacanth

- Category: General Science
  - When hit by electrons, a phosphor gives off electromagnetic energy in this form
  - Answer: light (or photons)

- Category: Lincoln Blogs
  - Secy. Chase just submitted this to me for the third time--guess what, pal. This time I'm accepting it
  - Answer: his resignation
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 majority of current question answering systems designed to answer factoid questions consist of three distinct components:
1) question analysis,
2) document or passage retrieval and finally
3) answer extraction.
Basic Architecture

Question & Topic Analysis → Question Decomposition → Hypothesis Generation → Hypothesis and Evidence Scoring → Synthesis → Final Confidence Merging and Ranking

- Question
- Multiple Interpretations
- 100s Sources
- 100s Possible Answers
- 1000's of Piece of Evidence
- 100,000's scores from simultaneous Text Analysis Algorithm

Answer & Confidence
Question Analysis

As the first component in a QA system it could easily be argued that question analysis is the most important part. Any mistakes made at this stage are likely to render useless any further processing of a question.
Determining the Expected Answer Type

Machine learning techniques to classify a question. We can train our system on thousands of tagged question corpus, provided by cognitive computation group at the department of computer science, University of Illinois at Urbana-Champaign to determine the expected answer.
Query Formation

The question analysis component of a QA system is usually responsible for formulating a query from a natural language question to maximise the performance of the IR engine used by the document retrieval component of the QA system.

We assume question itself is a valid IR query

We just remove stop words from the question
Database Access Schemata

- Who is the president of India?

Access Schemata –

Search <> for name <> biography.com <> person – president <> place - India
How Watson works: Step 1 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.
Document Retrieval

The text collection over which a QA system works tend to be so large that it is impossible to process whole of it to retrieve the answer. The task of the document retrieval module is to select a small set from the collection which can be practically handled in the later stages.
The Taj Mahal completed around 1648 is a mausoleum located in Agra, India, that was built under Mughal Emperor Shah Jahan in memory of his favourite wife, Mumtaz Mahal.
Pockets of structured and semi-structured knowledge

Entertainment
www.imdb.com

Places
www.cia.gov

Computers
zdnet.com

People
biography.com

Science
en.wikipedia.com

History
besthistorysites.net

Products
www.ebay.com
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...

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

Total Raw Content
Preprocessed Content

- 70 GB
- 500 GB
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.
Answer Extraction

Is responsible for ranking the sentences and giving a relative probability estimate to each one. It also registers the frequency of each individual phrase chunk marked by the NE recognizer for a given question class at a given rank.
Sense/Semantic similarity

- We use statistics to compute information content value.
- We assign a probability to a concept in taxonomy based on the occurrence of target concept in a given corpus.

We use Word Net as a sense/semantic dictionary

We obtain statistics of particular word from a large text corpus
Word Net - Synsets

screened

shady

dark

as in shady parts of garden

sense 1

as in shady character

sense 2

dishonest

sim(shady,dark)=0.656
sim(dark,dishonest)=0.0
The sentence as well as the query forms an ordered set of words. We then compute the sense network between every pair of words from query and sentence...

\[ \Gamma(w_i, q_j) = \xi_{i,j} \]

\( \xi_{i,j} \in [0, 1] \) is the value of sense/semantic similarity between \( w_i \in W \) and \( q_j \in Q \).

A sense network formed between a sentence and a query.
Exact Match Score

Given a sense network $\Gamma(w_i, q_j)$, we define the distance of a word $w_i$ as

$$d(w_i) = i$$

$$d(q_j) = j$$

Word with maximum sense similarity with query word $q_i$ is:

$$M(q_i) = w_j \mid j = \arg\max_j \xi_{j,i}$$

And the corresponding value of

$$\xi_{i,j} = V(q_i)$$

The exact match score is

$$E_{total} = \frac{\sum_i V(q_i)}{m}$$
Alignment Score

Let $T = \{\text{ordered set of } M(q_i) \, \forall \, i \in [1, m]\}$ in increasing order of $d(q)$. Function $\theta_i$ is the distance of $i^{\text{th}}$ element in $T$ then the alignment score is

$$K_{total} = \frac{\sum_{i=1}^{m-1} \text{sgn}(\theta_{i+1} - \theta_i)}{m-1}$$

An alignment score of 1.0 signifies perfect alignment while a score of -1.0 signifies reverse order of occurrence.
Total Score

We define the following coefficients

\[
\begin{align*}
\mu &= \text{noise penalty coefficient} \\
\lambda &= \text{sense similarity coefficient} \\
\psi &= \text{exact match coefficient} \\
\nu &= \text{order coefficient}
\end{align*}
\]

So the total score is a linear combination of individual scores

\[
\eta = \psi \times E_{total} + \lambda \times S_{total} + \mu \times \delta_{total} + \nu \times k_{total}
\]

We fine tune the values of these coefficients to get maximum accuracy.
Answer Confidence Score

We take top $t$ sentences and consider the plausible answers within them. If an answer appears with frequency $f$ in sentence ranked $r$ then that answer gets a confidence score -

$$C(ans) = \frac{1}{r} \left( 1 + \ln(f) \right)$$

all answers are sorted according to confidence score and top $9$ (=5 in our case) answers are returned along with corresponding sentence and URL.
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.
Automatic Learning for “Reading”

- **Volumes of Text**
- **Syntactic Frames**
  - subject
  - verb
  - object
- **Semantic Frames**
  - Inventors patent inventions (.8)
  - Officials Submit Resignations (.7)
  - People earn degrees at schools (0.9)
  - Fluid is a liquid (.6)
  - Liquid is a fluid (.5)
  - Vessels Sink (0.7)
  - People sink 8-balls (0.5) (in pool/0.8)
In cell division, mitosis splits the nucleus & cytokinesis splits this liquid cushioning the nucleus.

Many candidate answers (CAs) are generated from many different searches
Each possibility is evaluated according to different dimensions of evidence.
Just One piece of evidence is if the CA is of the right type. In this case a “liquid”.

Is(“Cytoplasm”, “liquid”) = 0.2
Is(“organelle”, “liquid”) = 0.1
Is(“vacuole”, “liquid”) = 0.2
Is(“plasma”, “liquid”) = 0.7

"Cytoplasm is a fluid surrounding the nucleus…"

Wordnet $\rightarrow$ Is_a(Fluid, Liquid) $\rightarrow$ ?

Learned $\rightarrow$ Is_a(Fluid, Liquid) $\rightarrow$ yes.
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.
In May 1898 Portugal celebrated the 400th anniversary of this explorer’s arrival in India.

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach.

The evidence is still not 100% certain.
A long, tiresome speech delivered by a frothy pie topping

Diatribe

Harangue

Meringue

Whipped Cream

Answer: Meringue  Harangue

Some Questions require Decomposition and Synthesis

Category: Edible Rhyme Time

Not Just for Fun
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.
Initial Question Formulated: “The name of this monetary unit comes from the word for "round"; earlier coins were often oval”

Watson performs question analysis, determines what is being asked.

It decides whether the question needs to be subdivided.
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...
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture

1. Initial Question
2. Question & Topic Analysis
3. Question Decomposition
4. Answer Sources
   - Primary Search
   - Candidate Answer Generation
5. Hypothesis Generation
6. Watson then uses algorithms to “score” each potential answer and assign a confidence to that answer...
7. Evidence Sources
   - Answer Scoring
   - Evidence Retrieval
   - Deep Evidence Scoring
8. Synthesis
   - If the question was decomposed, Watson brings together hypotheses from sub-parts

Watson uses Evidence Sources to validate it’s hypothesis and help score the potential answers.
DeepQA: the technology & architecture behind Watson: Massively Parallel Probabilistic Evidence-Based Architecture

Using models on the merged hypotheses, Watson can weigh evidence based on prior "experiences".

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
How we convert data into knowledge for Watson’s use

Three types of knowledge

- Domain Data (articles, books, documents)
- Training and test question sets w/answer keys
- NLP Resources (vocabularies, taxonomies, ontologies)

Converted to indices for search/passage lookup

Redirects extracted for disambiguation

Frame cuts generated with frequencies to determine likely context

Pseudo docs extracted for Candidate answer generation

Used to create logistic regression model that Watson uses for merging scores

Named entity detection, relationship detection algorithms

Custom slot grammar parsers, prolog rules for semantic analysis
Grouping features to produce Evidence Profiles

**Clue:** Chile shares its longest land border with this country.

---

**Positive Evidence**

Bolivia is more Popular due to a commonly discussed border dispute. But Watson learns that Argentina has better evidence.

---

Bolivia is more Popular due to a commonly discussed border dispute. But Watson learns that Argentina has better evidence.

---

© 2009 IBM Corporation
One Jeopardy! question can take **2 hours on a single 2.6Ghz Core Optimized & Scaled out on 2880-Core IBM workload optimized POWER7 HPC using UIMA-AS**, *Watson* answers in 2-6 seconds.
Watson: Precision, Confidence & Speed

• **Deep Analytics** – We achieved champion-levels of *Precision* and *Confidence* over a huge variety of expression.

- Emily Dickinson: 99%
- Walt Whitman: 60%
- Barnard: 10%

• **Speed** – By optimizing Watson’s computation for Jeopardy! on 2,880 POWER7 processing cores we went from 2 hours per question on a single CPU to an average of just 3 seconds – fast enough to compete with the best.

• **Results** – in 55 real-time sparring against former Tournament of Champion Players last year, Watson put on a very competitive performance, winning 71%. In the final Exhibition Match against Ken Jennings and Brad Rutter, Watson won!
Watson-enabled patient-centered healthcare solutions

- Patient
- Lay Caregiver...
- PA...
- Nurse Practitioner
- Physician

- Consumer Portal
- Patient Inquiry
- Coding Automation
- Patient Workup
- Differential Diagnosis
- Specialty Diagnosis & Treatment Options
- Treatment Options
- Treatment Authorization
- Treatment Protocol Analysis
- Care Consideration Analysis
- Population Analysis & Care Mgmt
- Second Opinion
- On-going Treatment
- Specialty Research
- Genomic-based Analysis
- Longitudinal Patient Electronic Health Information
- What's New?
- Specialty Diagnosis & Treatment Options
- Caregiver Education
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
Our Study Path Forward for “Natural Language for Communication”

• Groundwork:
  – Review of probability: Ch. 13
  – Probabilistic reasoning over time: Ch. 15.1-15.3
  – Language models: Ch. 22.1

• Natural language for communication: Ch. 23