In February 2011, IBM Watson defeated Brad Rutter and Ken Jennings in the Jeopardy! Challenge.
Today, Watson is a **cognitive system** that enables a new partnership between people and computers that enhances and scales human expertise by providing a more **natural** relationship between the human and the computer.
What is Cognitive Computing?

• Cognitive systems are able to learn their behavior through education
• That support forms of expression that are more natural for human interaction
• Whose primary value is their expertise
• That continue to evolve as they experience new information, new scenarios, and new responses
Technical Publications on Watson NLP

This talk presents technical features published in 17 articles the 2012 *IBM Journal of Research and Development* (available to IEEE subscribers) plus IBM documentation.
Question Answering

One part of Watson’s cognitive computing platform is **Question Answering**. The main objective of QA is to analyze natural language questions and present **concise answers with supporting evidence**, rather than a list of possibly relevant documents like internet search engines.
This talk will have two parts

**Part 1 – NLP**
I will describe some of the natural language processing components.

**Part 2 – QA Pipeline**
I will describe three of the basic stages of IBM Watson’s Question Answering pipeline.
Part 1 – NLP
The NLP components that help make the QA pipeline work include

- Full syntactic parse
- Entity and relationship extraction
- Semantic tagging
- Co-reference
- Automatic frame discovery - PRISMATIC
- Many others

This talk will discuss how sophisticated linguistic resources allow Watson to achieve true question answering functionality.
NLP in Watson

Watson uses many NLP tools and techniques, including these novel, specialized utilities:

- Slot Grammar Parser
- Predicate Argument Structure
- PRISMATIC
Slot Grammar

• The linguistic soul of Watson.
• Slot Grammar is a rule-based dependency parser.
• Produces a complete, sentence level structural representation of syntax and semantics.
• Uses large and extensible lexicons.
• Lexicons provide domain adaptation capability

Slot Grammar

Lexicons tell Slot Grammar what to do when it consumes words in order to build a full sentence-level structure.

Lexicons define

• Words (and MWEs)
• Parts of Speech: verb, noun, preposition
• Argument slots (subject, object, complement)
• Semantic constraints: proper noun, motion verb
• Multiple-senses
## Slot Grammar

### Toy Lexicon

<table>
<thead>
<tr>
<th>word</th>
<th>POS</th>
<th>required slots (options)</th>
<th>semantic types</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>hit</strong></td>
<td>verb</td>
<td><code>1-SUB hit 2-OBJ (with 3-IndOBJ)</code></td>
<td>violence</td>
</tr>
<tr>
<td><strong>hand</strong></td>
<td>verb</td>
<td><code>1-SUB handed 2-OBJ to 3-IndOBJ</code></td>
<td>giving</td>
</tr>
<tr>
<td><strong>department</strong></td>
<td>noun</td>
<td><code>department (of 1-COMP)</code></td>
<td>institution</td>
</tr>
<tr>
<td><strong>bank</strong></td>
<td>noun</td>
<td>(&quot;national&quot;</td>
<td>&quot;federal&quot;)</td>
</tr>
<tr>
<td><strong>bank</strong></td>
<td>noun</td>
<td>(&quot;river&quot;)</td>
<td>geographic</td>
</tr>
<tr>
<td><strong>bank</strong></td>
<td>verb</td>
<td><code>bank (&quot;the nine ball&quot;)</code></td>
<td>skill</td>
</tr>
</tbody>
</table>
Slot Grammar - Output

Figure 1

ESG parse of the Jeopardy! question “Chandeliers look great but nowadays do not usually use these items from which their name is derived.”

Predicate Argument Structure - PAS

A full SG parse can hurt text similarity discovery because it is too syntactically rigid. The PAS builder provides simplification and abstraction of the ESG parse by removing some details. E.g., PAS prunes auxiliary nodes and simplifies POS tags in order to find similarity. The following five sentences all get the same PAS structure.

\begin{itemize}
  \item Edison invented the phonograph in 1877.
  \item Edison invented a phonograph in 1877.
  \item Phonographs were invented by Edison in 1877.
  \item Edison was busy inventing the phonograph in 1877.
  \item The phonograph was invented by Edison in 1877.
\end{itemize}
SG Parse vs. PAS

A simple comparison. PAS removes auxiliary verbs like “be”, and semantic tags like “artf” (artifact) and “langunit”.

PRISMATIC

- Automatic frame induction
- Large scale, lexicalized, syntactic relation resource
- After parsing large corpus, PRISMATIC discovers common lexical-grammatical patterns associated with words.
- Patterns like S-V-O, S-V-P-O and S-V-O-IO


PRISMATIC

• Slot grammar parses a large corpus.
• PRISMATIC discovers that objects of the verb “annex” are statistically likely to be geopolitical entities.
• Watson does not know anything about the string “Piedmont”.
• Given the phrase “Napoleon annexed Piedmont”, Watson can infer that *Piedmont* is a geopolitical entity.
Part 2 – QA Pipeline
Using NLP for QA

Three Stages (there are more...)
- Question Analysis
- Hypothesis Generation
- Semantic Types
Question Analysis

• **Focus** – The segment of a question in which a substituted answer produces a statement.

• **Lexical Answer Type (LAT)** – The type of answer sought by a question. Detection is done via Prolog and statistical learning.

• **Question Class** – Factoid, Definition, List, etc
Question Analysis

Watson uses the parse to identify useful LATs.

**Focus**

**LAT**

The focus “who” is the subject of the verb “is” and in this case the LAT is the object of that verb, “inventor”.

Who is the inventor of the phonograph and the light bulb?

**Focus**

**LAT**

In this case, the focus word “which” is the determiner of the LAT noun “novel, which is the headword of an NP attached to the wh-word “which”.

Which Peter Benchley novel is about a giant squid that menaces the coast of Bermuda?
Hypothesis Generation

• The process of producing possible answers to a given question
• Watson uses the knowledge inherent in human-generated text and associated metadata, as well as syntactic and lexical cues in search results, to identify salient concepts from text and hypothesize them as candidate answers.
• PRISMATIC is a crucial component of hypothesis Generation.

Hypothesis Generation

"Which country singer was sent to jail for robbery and in 1972 was pardoned by Ronald Reagan?"

- This wiki page is linguistically similar to the question.
- LAT “singer” is of type PERSON (Slot Grammar).
- Title is a person name (Named Entity tagging).
→ Hypothesize that the wiki title text is a possible answer.

Hypothesis: Merle Haggard

Merle Ronald Haggard (born April 6, 1937) is an American country and Western song writer, singer, guitarist, fiddler, and instrumentalist. Along with Buck Owens, Haggard and his band the Strangers created the Bakersfield sound, which is characterized by the unique twang of a Fender Telecaster and the unique mix with the traditional country steel guitar sound, new vocal harmony styles in which the words are minimal, and a rough edge not heard on the more polished Nashville Sound recordings of the same era.

In the 1970s, Haggard was aligned with the growing outlaw country movement, and has continued to release successful albums through the 1980s and into the 2000s. In 1994, he was inducted into the Country Music Hall of Fame. In 1997, he was inducted into the Oklahoma Music Hall of Fame.

Married and plagued by financial issues, he was arrested in 1957 shortly after he tried to rob a Bakersfield roadhouse. He was sent to Bakersfield Jail and was later transferred after an escape attempt to San Quentin Prison, on February 21, 1958. While in prison, Haggard discovered that his wife was expecting a child from another man, which pressed him psychologically. He was fired from a series of prison jobs, and planned to escape along with another inmate nicknamed “Rabbit”. Haggard was convinced not to escape by fellow inmates. Haggard started to run a gambling and brewing racket with his cellmate. After he was caught drunk, he was sent for a week to solitary confinement where he encountered Caryl Chessman, an author and death row inmate. Meanwhile, “Rabbit” had successfully escaped, only to shoot a police officer and return to San Quentin for execution.

Haggard soon earned a high school equivalency diploma and kept a steady job in the prison's textile plant, while also playing for the prison’s country music band, attributing a 1958 performance by Johnny Cash at the prison as his main inspiration to join it. Upon his release in 1960, Haggard said it took about four months to get used to being out of the penitentiary and that, at times, he actually wanted to go back in. He said it was the loneliest he had ever felt.

According to Rolling Stone, "In 1972, then—California governor Ronald Reagan expunged Haggard's criminal record, granting him a full pardon."
Semantic Types

Human language is remarkably rich when it comes to assigning types; nearly any word can be used as a type, particularly in some questions.

- Invented in the 1500s to speed up the game, this maneuver involves two pieces of the same color. (Answer: Castling)
- The first known airmail service took place in Paris in 1870 by this conveyance. (Answer: hot-air balloon)
- In 2003, this Oriole first sacker was elected to the Baseball Hall of Fame. (Answer: Eddie Murray)
- Freddy Krueger was introduced in this 1984 scarefest. (Answer: A Nightmare on Elm Street)
- When hit by electrons, a phosphor gives off electromagnetic energy in this form. (Answer: light)
Semantic Types
So many types, so little time.
Semantic Types
Watson is open and flexible about types. Instead of finding candidates of the right type, we want to find and judge whether each one is of the right type by examining it in context with the answer type from the question. *Find anything in Watson knowledge that suggests Candidate and LAT are the same type.*

---

**Figure 2**
High-level architecture of the TyCor component involving four core steps. For example, in YAGO TyCor, given candidate “difficulty swallowing” and lexical answer type (LAT) “manifestation”, EDM maps candidate to DBpedia entity “Dysphagia”; TR obtains WordNet type “Symptom” for the DBpedia instance; PDM maps LAT to WordNet concept “Condition”; and the final TR step finds a hyponymy relation between “Symptom” and “Condition” producing a positive TyCor score.
Open Source and Watson

IBM has committed to open source with contributions to more than 120 projects. IBM Watson also utilizes open source.

**WordNet** - a large lexical database of English.

**UIMA** - Unstructured Information Management Applications. UIMA is an Apache-licensed open source framework that enables applications to be decomposed into NLP components.

**Lucene** - Apache Lucene is a free open source information retrieval software library. It is supported by the Apache Software Foundation and is released under the Apache Software License.

**Dbpedia** - DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia.
References


Thanks!

Questions?