Artificial intelligence meets Big and Open Data

Nobuyuki Igata
Research Manager
Knowledge Platforms Laboratory
Social Innovation Laboratories
Fujitsu Laboratories Ltd. (Japan)
Contents

- Why Artificial Intelligence is now again

- New wave in World Wide Web
  - Web of data: Linked Data
Why Artificial intelligence is now again
Overview of AI trend

**Evolution of AI**

1980

**Expert System**

1990

**DeepBlue**
512 Core
Millions Books of data

2000

**DeepQA**
2,880 Core
Millions Books of data

2010

**Deep Learning**
16,000 Core
10 Millions images

2011

2012

**How many Web site?**

<table>
<thead>
<tr>
<th>Years</th>
<th>1995</th>
<th>2006</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Millions site</td>
<td>18,000</td>
<td>100M</td>
<td>1B</td>
</tr>
</tbody>
</table>

**Performance of ICT**

- CPU
- Storage
- Network

- Exponentially increased

Copyright 2014 FUJITSU
Deep Learning

- Image recognition by Google 2012
  - 16,000 CPU Core
  - Deep Learning Algorithm
    - Based on multi-layer Neural network
  - 10 Millions images from YouTube
  - Autonomously Learning for 1 week
    - (so-called Unsupervised Learning)
  - Finally the system can distinct face of Cats and Human
  - Automatic feature extraction

Source: Quoc V. Le, etc. “Building high-level features using large-scale unsupervised learning”, ICML2012
Computer Shōgi

Shōgi

- Japanese Traditional Board Game
- Similar rules with Chess

Difficulties compared with Chess

<table>
<thead>
<tr>
<th></th>
<th>Chess</th>
<th>Shōgi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Board Size</td>
<td>64: 8×8</td>
<td>81: 9×9</td>
</tr>
<tr>
<td>Num of Pieces</td>
<td>32</td>
<td>40</td>
</tr>
<tr>
<td>Different Pieces</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Legal Positions</td>
<td>$10^{47}$</td>
<td>$10^{71}$</td>
</tr>
<tr>
<td>Possible Games</td>
<td>$10^{123}$</td>
<td>$10^{226}$</td>
</tr>
</tbody>
</table>

Develop a Strong Computer Shōgi from Data

Past Approach

Human programmer
Strong Shōgi Player

Human extract features from his experience

500 features

Amateur-level

Machine Learning based Approach

Past (Professional)
Game results

Automatic Feature Extraction from past data

100M features

Top-level professional

Past (Professional)
Game results

Value of Pieces

Position of Pieces

87 points, 569 points

Top-level professional
DeepQA: Question Answering

- IBM Watson won Quiz Champion on *Jeopardy!* 2011
  - 2,880 CPU Core
  - Millions Books of data
  - Combined several techniques
    - NLP, IR, KR, ML, Reasoning

Tōdai Robot Project in Japan

- "Can a robot pass entrance exam of the University of Tokyo (Todai) ? ”
- Grand Challenge by National Institute of Informatics, Japan (2011 – 2021)
  - An open platform to evaluate various AI technologies in a real-world situation
  - Collaboration of a variety of AI technologies
- Fujitsu Labs joins to this project especially Mathematical exam
  - Need Reasoning mechanism like Human thoughts
"Find the radius of a circle with area $4\pi$."

$\exists r \left[ \forall C. \text{circle}(C) \land \text{area}(C) = 4\pi \rightarrow \text{radius}(C) = r \right]$

$\exists r \left[ \forall s. \pi s^2 = 4\pi \land s > 0 \rightarrow s = r \right]$

$r = 2$

Math Knowledge Base

- Quantifier Elimination
- Groebner Basis
- Theorem Prover
- etc.

Math Solver System

- Parsing
- Anaphora resolution
- Discourse analysis

Question

Language Understanding

Logical Form

Formula Rewriting

Solver Input

Solver algorithms

Answer

Math Lexicon (dictionary)
座標平面上に\{4点 O(0, 0), A(3, 0), B(3, 2), C(0, 2)\}を頂点とする長方形がある。\{辺AB, BC上に\}\{動点 P(3, p), Q(q, 2)\}を\{OP = PQ\}をみたすようにとる}。このとき、\triangle OPQ の面積を S とする。

(1) p を q を用いて表せ。次に、q のとりうる値の範囲を求めよ。
(2) S の最小値を求めよ。
“Find the radius of a circle with area $4\pi$."

$\exists r \left[ \forall C. \text{circle}(C) \land \text{area}(C) = 4\pi \rightarrow \text{radius}(C) = r \right]$

$\exists r [ \forall s. \pi s^2 = 4\pi \land s > 0 \rightarrow s = r ]$

$r = 2$

How to create Knowledge Base

Math Knowledge Base

Math Lexicon (dictionary)

- Parsing
- Anaphora resolution
- Discourse analysis

- Quantifier Elimination
- Groebner Basis
- Theorem Prover
- etc.
Background: Open Data

- G8 (Fr, US, UK, DE, JP, IT, CA, RU)
  - Open Data Charter released in June 2013
  - Identified 14 high-value categories
  - Commit to releasing data by Dec. 2015
  - Provide data in Machine-readable format

- Over 40 countries in the world have websites for data disclosure
- Japan cabinet office defined the E-government open data strategy

![Global Open Data Index](http://global.census.okfn.org/)

From [http://global.census.okfn.org/](http://global.census.okfn.org/) by Open Knowledge Foundation
What is Machine-readable format?

5 Star Open Data
By Tim.B Lee

- Available on the Web under an open license e.g., PDF
- Available as structured data e.g., Excel
- Use non-proprietary formats e.g., CSV
- Use URIs so that others can point at your stuff
- Link your data to other people’s data to provide context e.g., Linked Data

Source: http://5stardata.info/
25 years ago, World Wide Web appeared in the internet. Web has a link structure called HyperText and provided the mechanism to disclose documents that had been stored in individual computers.
Web of Documents

From the middle of 90’s, many kinds of information had been open to the public on Web. Web brought a dramatic revolution to the method of information accesses.
Web of Documents

However, traditional web is optimized to “Human read it and understand it”, individual data still remains inside each document.

Human-readable
And then Linked Data appears to solve this issue. Linked data can release individual data from each document. And data can have a link to other data like web document.
In Linked Data, the semantics or the relationship of data are represented by “Link”. Computer can easily understand the semantics of data. This is the reason why Linked Data is Machine-readable.
From 2007, dataset described in Linked Data format has been released on Web. Currently these dataset make a big data network on Web. These dataset and data network are called “Linked Open Data (LOD)”.
Activities of Fujitsu Labs

Data Conversion

Federated different datasets

LOD Search Engine

Retrieve LOD in the world

http://lod4all.net/
Current Status and Issues of LOD

- Many web sites have published their own data separately

**Problem 1**
I don’t know which site has the data I want

**Problem 2**
Requires complex processing in the application to do the processing that combines data from multiple sites

**Problem 3**
Data in a site without SPARQL endpoint is not searchable

User

Application

Internet

LOD Site-A

LOD Site-B
LOD Search Engine by Fujitsu Labs

- Store LOD from around the world (several tens of billion data items), and provide a high-speed search

Solution 1
Use data without being aware of what site it is on

Solution 2
Applications need not to implement complex processes!

Solution 3
Can even search sites without search functions!

User Application

Search UI Standard API
LOD Search Engine

Collect Collect Collect
LOD Site-A LOD Site-B
LOD Search Engine by Fujitsu Labs

http://lod4all.net

Demo Video

LOD4ALL Search

or

http://lod4all.net
Heterogeneous Financial Data

- How to establish interoperability between heterogeneous data?
  - No-links, data silos
  - Format heterogeneity (XML, schemas, APIs, Web crawling, CSV files, etc.)
Use case: Financial Domain

- Financial Reports, etc.
- LEI: Legal Enterprise Identifier
- Market Data (Stock Price)

Analysts

Application

Dashboard

Standard API

LOD Search Engine

Data Conversion

LOD Crawling

Public

Linked Open Data (LOD)

NY Times

Crunchbase

Public

Public

Crunchbase

DBpedia

Crunchbase

DBpedia

NY Times

Public
Use case: Financial Domain
LOD for all: Realising Data-utilization Services

Go to Demo#A14!

LOD for all

Unified access to LOD across the world in a batch

http://lod4all.net/

LOD Search Engine

Financial

Healthcare

Open Government

Tourism
Future of Artificial Intelligence with Linked Data

- **Hyper-connected World (IoT)**
  - Billions devices connected to Internet
  - Deep Learning has the potential to realise *Artificial Senses* like human

- **Linked Data/LOD**
  - Linked data is not just data format but also *Description Logic* based
    - Todai Robot solves questions by using family logic (First-order predicate Logic)
  - LOD has the potential to be used as **Web Scale Knowledge Base** for *Artificial Thought*

**Billions Devices Connected**

**Huge Graph High Complexity**

**Web scale Artificial intelligence**

I can understand First-order predicate logic