Knowledge Extraction in Text

Marco Ponza
Who am I? Last-year PhD student...

Advanced Algorithms & Applications Lab
supervisor: Prof. Paolo Ferragina

1. Data Compression
2. Searching & Mining Web & Social Media
3. Natural Language Understanding
Natural Language Understanding

Easy for humans ✔

Hard for machines ✗

But, machines need today to access, read and understand information stored in very large data archives

...and this will get to be more and more crucial with Conversational AI systems!
Machines represent texts by their *(possibly, ambiguous)* words

*Leonardo is the scientist who painted *Mona Lisa*
Machines represent texts by their (possibly, ambiguous) words

Leonardo is the scientist who painted Mona Lisa
Machines represent texts by their (possibly, ambiguous) words

Leonardo is the scientist who painted Mona Lisa
Knowledge Graph
May 2012

https://www.blog.google/products/search/introducing-knowledge-graph-things-not
Leonardo is the scientist who painted Mona Lisa.
Understanding the Text by Entities, not Strings

Leonardo is the scientist who painted Mona Lisa

Science

Italy

Leonardo Da Vinci

Mona Lisa (painting)

Leonardo Da Vinci

Mona Lisa (painting)

Cartography

Art

Renaissance

Since 2010: Efficient & effective solutions for this problem!

A³Lab

TAG-ME

Google Faculty Research Awards
Two Editions: 2010 & 2013
Don't look for presidential announcements from Senators Barack Obama and Hillary Rodham Clinton anytime soon, but stay tuned.

At least that is the word from their associates. Mr. Obama, Democrat of Illinois, is not likely to say whether he intends to seek the party's presidential nomination until after President Bush's State of the Union address on Jan. 23. As he walked out of the Capitol on a recent afternoon, Mr. Obama only smiled when asked about his timing. Then, he rushed to change the subject.

Initially, Mr. Obama said he intended to announce his decision after returning from a holiday vacation in Hawaii, where he was visiting his grandmother and other relatives. Now, several people close to the senator say, he needs a little more time to make up his mind.

Still, Mr. Obama has been busy telephoning crucial Democrats in Iowa, New Hampshire and other states. There is, of course, only one reason for him to be making such inquiries.

Last week on Capitol Hill, Mr. Obama bumped into Ethel Kennedy, who has been a big admirer. When asked about him, she said, "He can't run soon enough."

Mrs. Clinton, meanwhile, plans to announce her decision in the next several weeks, her advisers say. According to several Democrats who have spoken to her, as well as advisers, Mrs. Clinton has given every indication that she is running, short of saying so, and no signals that she is not.

She is making phone calls to Democratic officials, labor leaders and supporters in early nominating states. And she continues to talk to possible consultants and donors, yet she has not made any travel plans to kick off a campaign.
Understanding the Text by Salient Entities

Entity Salience Problem
Relevant vs Non-Relevant Entities

...and more and more entities!
Understanding the Text by **Salient Entities**

- **Entity Salience Problem**
  Relevant vs Non- Relevant Entities

- **Our Solution**
  - Improvements of +12% wrt CMU/Google system
  - Published at [NLdb 2017](#)

*...and more and more entities!*
Understanding the Text by **Salient Entities**

- **How? Applying Graph Theory and Algorithms!**
  
  - Used to draw new features
  - Classify entities into salient/non-salient via ML
  
- **Problem: How can we weight the edges?**
Understanding the Text by **Extraction of Facts**

- How to **enrich** the Knowledge Graph with **information** that comes from a **text**?
  1. Identify **salient entities**
  2. Extract **facts** connecting them

---

*Leonardo is the scientist who painted Mona Lisa*

**Facts**

- (“Leonardo”, “is”, “scientist”)
- (“Leonardo”, “painted”, “Mona Lisa”)

---

**max planck institut informatik**
Expert Finding & Profiling

- ~1.5K Authors
- ~65K Documents (papers’ abstracts)
- ~35K Research Topics
- More than 1K queries and ~2K profiles view in few months
- Currently used by UniPi’s Technology Transfer Office
Select the range of years to analyze.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Entity</th>
<th>Count</th>
<th>Doc. count</th>
<th>Years</th>
<th>Wiser score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Recurrent neural network</td>
<td>26</td>
<td>15</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Artificial neural network</td>
<td>44</td>
<td>24</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Tree (data structure)</td>
<td>24</td>
<td>17</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Machine learning</td>
<td>16</td>
<td>14</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Recursive neural network</td>
<td>16</td>
<td>16</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Quantitative structure-activity</td>
<td>30</td>
<td>16</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Echo state network</td>
<td>14</td>
<td>13</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Mathematical model</td>
<td>39</td>
<td>28</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Prediction</td>
<td>19</td>
<td>17</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Generative model</td>
<td>9</td>
<td>7</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Group representation</td>
<td>22</td>
<td>14</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Empiricism</td>
<td>15</td>
<td>15</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Data model</td>
<td>8</td>
<td>8</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Data set</td>
<td>20</td>
<td>14</td>
<td>1998-2018</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Reservoir computing</td>
<td>8</td>
<td>8</td>
<td>1998-2018</td>
<td></td>
</tr>
</tbody>
</table>
Future Directions

The Alexa Prize
Over $3.5 Million to Advance Conversational Artificial Intelligence
December 2017 - November 2018

Moving the paradigms: from text to conversations!
https://sobigdata.d4science.org/web/tagme

The Entity Linking tools by Acube lab

Welcome to the Tagme Virtual Research Environment. From here, you can access all Entity Linking tools provided by the Acube lab at the University of Pisa.

- **TAG-ME**: Entity linker, ideal for annotating noisy text. Available languages: en, de, it
- **SWAT**: Entity linker, ideal for annotating well-formed text. More accurate than Tagme, but still experimental. Available languages: en
- **SMAPH**: Entity linker for web search queries. Available languages: en
- **SWAT**: Entity Salience service: assigns a relevance score to the entities mentioned by a document. Available languages: en

Thanks!