INTRODUCTION

• Information is typically available both in unstructured and structured form

• Deep and Large scale NLP now enables to bridge the two "world"

• Development of frameworks for integrating unstructured and structured content only partially investigated
KnowledgeStore

• A scalable, fault-tolerant, and Semantic Web grounded storage system to jointly store, manage, retrieve, and query, both structured and unstructured data.
Among a collection of news articles, a user is interested in retrieving all 2014 articles reporting statements of a 20th century US president where he is positively mentioned as “commander-in-chief”.
• Among a collection of news articles, a user is interested in retrieving all 2014 articles reporting statements of a 20th century US president where he is positively mentioned as “commander-in-chief”.
"Stripes," Dole, 1996

He (Bill Clinton) quickly became the first civilian commander-in-chief to salute his marine guards while entering or exiting an aircraft.
Exploitation

• Enhanced applications (e.g., decision support systems)

• Developing, debugging, training, and evaluating NLP and knowledge processing tasks

• Reasoning on Extracted Information (e.g., on Events)

• Text Exploration
OUTLINE

• A walk through the KnowledgeStore
• The KnowledgeStore “live”
• The KnowledgeStore in NewsReader
• Next Challenges
KnowledgeStore

Functional View

Representation
- Hadoop HDFS
  - multiple processes

Resource
- HBase
  - multiple nodes

Mention
- Virtuoso
  - single process

Entity
- Axiom

KnowledgeStore

Populators
- for resources and RDF
- Text
- RDF

NLP Processors
- single / cross-document

Applications

UI

API
Knowledge Store

Data Model

Resources
- arbitrary text documents, e.g., news and associated annotation files
- both digital representation (file) and associated metadata attributes

Mentions
- snippets of resources denoting something of interest - entity / axiom
- sign vs referent distinction
- described by linguistic metadata attributes

Entities
- persons, organizations, events, ...
- described by logical axioms (~triples)
- axiom validity restricted to a specific context (e.g., time span, attribution pair)

- It’s Flexible
- It’s an OWL 2 Ontology
Example: Data Model For Event Extraction from News

Namespaces:
- nwr: <http://dkmfbk.eu/ontologies/newsreader#> (default, omitted)
- ks: <http://dkmfbk.eu/ontologies/knowledgestore#>
- dcterms: <http://purl.org/dc/terms/>
- gaf: <http://groundedannotationframework.org/gaf#>
- nie: <http://www.semanticdesktopen ontology/ontologies/2007/01/19/nie#>
- nfo: <http://www.semanticdesk topten ontologies/2007/03/22/nfo#>
- nif: <http://nl2rdf.lod2.eu/schema/string/>
- owtime: <http://www.w3.org/2006/time#>
KnowledgeStore

Architectural View

- Any application (HTTP access to the KS, possibly exploiting SPARQL client libraries)
- Java applications
  - KnowledgeStore Java client

Client-side

- Zookeeper (multi. nodes)
  - distributed synchronization
- Hadoop HDFS (name & data nodes)
- HBase (multiple server nodes)

Server-side

- KnowledgeStore Frontend Server
  - SPARQL endpoint
  - CRUD endpoint
- Virtuoso (single node)
  - Entity
  - Axiom
  - Context
- OMID (single node)
  - transaction manager
• The KnowledgeStore **User Interface**

• lookup: given the URI of an object (i.e., resource, mention, entity), retrieves all the content about it

• SPARQL: run arbitrary queries against the SPARQL endpoint
Knowledge Store

NewsReader

recording history
by processing massive streams of daily news

ICT 316404
FP7-ICT-2011-8
www.newsreader-project.eu
Jan 2013 - Dec 2015
Dataset used in the Demo (NewsReader Project):

- **Domain:** Global Automotive Industry

- **1.3M news documents** (2003-2013), provided by LexisNexis (www.lexisnexis.nl)

- 1.3M NLP annotation files (NAF format) obtained processing the news (NewsReader NLP Pipeline)

- **205M mentions** of events, persons, organisations, locations, time expressions...

- **535M of RDF triples** about events, persons, organisations, locations, time expressions...
  - 439M extracted from text
  - 96M coming from selected background knowledge (DBpedia)
<table>
<thead>
<tr>
<th>Domain</th>
<th>WikiNews</th>
<th>Cars (Ver. 1)</th>
<th>FIFA WorldCup</th>
<th>Cars (Ver. 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource Providers</td>
<td>General News</td>
<td>Automotive Industry</td>
<td>Sport, Football</td>
<td>Automotive Industry</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Resources</td>
<td>18,510</td>
<td>63,635</td>
<td>212,258</td>
<td>1,259,748</td>
</tr>
<tr>
<td>Mentions</td>
<td>2,629,176</td>
<td>9,110,683</td>
<td>76,165,114</td>
<td>205,114,711</td>
</tr>
<tr>
<td>Entities</td>
<td>951,879</td>
<td>2,212,691</td>
<td>10,246,338</td>
<td>27,123,724</td>
</tr>
<tr>
<td>Events</td>
<td>624,439</td>
<td>1,783,991</td>
<td>9,387,356</td>
<td>25,156,574</td>
</tr>
<tr>
<td>Persons</td>
<td>94,731</td>
<td>199,999</td>
<td>403,021</td>
<td>729,797</td>
</tr>
<tr>
<td>Organizations</td>
<td>101,754</td>
<td>187,842</td>
<td>431,232</td>
<td>947,262</td>
</tr>
<tr>
<td>Locations</td>
<td>130,955</td>
<td>40,859</td>
<td>24,729</td>
<td>290,091</td>
</tr>
<tr>
<td>Axioms (Triples)</td>
<td>105,610,963</td>
<td>316,034,616</td>
<td>240,731,408</td>
<td>535,011,673</td>
</tr>
<tr>
<td>from Mentions</td>
<td>9,700,585</td>
<td>46,359,300</td>
<td>136,135,841</td>
<td>439,101,295</td>
</tr>
<tr>
<td>from Background Knowledge</td>
<td>95,910,378</td>
<td>269,675,316</td>
<td>104,595,567</td>
<td>95,910,378</td>
</tr>
<tr>
<td>distilled from:</td>
<td>DBpedia 2014 (EN)</td>
<td>DBpedia 3.9 (EN,NL,ES,IT)</td>
<td>DBpedia 3.9 (EN)</td>
<td>DBpedia 2014 (EN)</td>
</tr>
<tr>
<td>Total Disk Space (GB)</td>
<td>17.64</td>
<td>30.67</td>
<td>82.48</td>
<td>260.20</td>
</tr>
<tr>
<td>Resource Layer</td>
<td>1.25</td>
<td>3.10</td>
<td>16.55</td>
<td>108.27</td>
</tr>
<tr>
<td>Mention Layer</td>
<td>1.49</td>
<td>4.77</td>
<td>41.72</td>
<td>112.00</td>
</tr>
<tr>
<td>Entity Layer</td>
<td>14.90</td>
<td>22.80</td>
<td>24.21</td>
<td>39.93</td>
</tr>
<tr>
<td>Approx. Population Total Time (hrs)</td>
<td>2</td>
<td>30</td>
<td>56</td>
<td>160</td>
</tr>
<tr>
<td>Approx. Rate (resources/hour)</td>
<td>9,300</td>
<td>2,250</td>
<td>4,000</td>
<td>7,800</td>
</tr>
</tbody>
</table>
KnowledgeStore in NewsReader

Decision Making on top of the KnowledgeStore
KnowledgeStore in

Exploited in Three Hack Day Events
Exploited in Three Hack Day Events

- Capable of handling:
  - large number of requests (>110K)
  - multiple concurrent requests (40 requests/sec.)
  - low response time (30-214ms)
Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)

- Example: “Yesterday, Chrysler hired Jim Press to lead its sales and marketing”
Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)

- Example: “Yesterday, Chrysler hired Jim Press to lead its sales and marketing”

At time: 2007/09/07

- **Event:** hire
- **Event type:** ESO:JoiningAnOrganization
- **Event roles:**
  - ESO:employer dbpedia:Chrysler
  - ESO:employee dbpedia:Jim_Press
Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)

- Example: “Yesterday, Chrysler hired Jim Press to lead its sales and marketing”

**Event:** hire
**Event type:** ESO:JoiningAnOrganization
**Event roles:**
- ESO:employer dbpedia:Chrysler
- ESO:employee dbpedia:Jim_Press

At time: 2007/09/07

**pre-Situation**
- dbpedia:Jim_Press
- ESO:notEmployedAt dbpedia:Chrysler
Inferring Knowledge **not Explicitly Mentioned** in Text (powered by Event Situation Ontology - ESO)

- Example: “Yesterday, Chrysler hired Jim Press to lead its sales and marketing”

<table>
<thead>
<tr>
<th>Event:</th>
<th>hire</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event type:</td>
<td>ESO:JoiningAnOrganization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pre-Situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia:Jim_Press, ESO:notEmployedAt, dbpedia:Chrysler</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>post-Situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia:Jim_Press, ESO:employedAt, dbpedia:Chrysler</td>
</tr>
</tbody>
</table>

At time: 2007/09/07
KnowledgeStore

Reasoning on Events

- Applied on the 1.3M global automotive industry news

- Extremely fast: 1,333s (~22m) to process ~500M triples

- Triggered 2M new triples (i.e., not in the text), organised in 397,885 situations

- 255,470 events have at least a pre/post/during situation:
  - 71.2% of the events having at least two distinct roles
KnowledgeStore

Looking ahead to the future...
Beyond “Asserted” Knowledge…

• What knowledge can be additionally inferred from what mentioned in text?
What knowledge can be additionally inferred from what mentioned in text?
Beyond “Asserted” Knowledge...

- What knowledge can be additionally inferred from what mentioned in text?

Event: hire
Event type: ESO:JoiningAnOrganization
Event roles: ESO:employer B, ESO:employee A
At time: 2007/09/07

Event: fire
Event type: ESO:LeavingAnOrganization
Event roles: ESO:employer C, ESO:employee A
At time: 2011/08/05
Beyond “Asserted” Knowledge…

- What knowledge can be additionally inferred from what mentioned in text?
What knowledge can be additionally inferred from what mentioned in text?
Beyond “Asserted” Knowledge…

• What knowledge can be additionally inferred from what mentioned in text?
Beyond “Asserted” Knowledge…

What knowledge can be additionally inferred from what mentioned in text?
What knowledge can be additionally inferred from what mentioned in text?
What knowledge can be additionally inferred from what mentioned in text?

Can we infer that some event took place?

What is the most probable one (if any)?
Knowledge Crystallisation

- When can a fact **automatically extracted** from information extraction tools be considered as “**Background Knowledge**”?

- Aspects to be considered:
  - “**cleaning**” of the data
  - **consistency** of event
    - “John” has been fired today from company “Alpha AGf”, but the KnowledgeStore contains the fact that “Alpha AGf” was closed 4 years ago
  - event **matching** and **integration**
    - augment background knowledge entities with facts extracted from the pipeline

- How To Tackle This? Combining **Statistical** and **Crisp** Reasoning?
Integrating knowledge extracted from **text** (news, twits,…), **pictures**, **movies**…

- E.g., retrieve all documents and media where Bill Clinton makes statements about the Army
Integrating knowledge extracted from **text** (news, twits, …), **pictures**, **movies** …

- E.g., retrieve all documents and media where Bill Clinton makes statements about the Army

"Stripes," Dole, 1996

He (Bill Clinton) quickly became the first civilian commander-in-chief to salute his marine guards while entering or exiting an aircraft.
• Integrating knowledge extracted from **text** (news, twits,…), **pictures, movies**…

• E.g., retrieve all documents and media where Bill Clinton makes statements about the Army

"Stripes," Dole, 1996

He (Bill Clinton) quickly became the first civilian commander-in-chief to salute his marine guards while entering or exiting an aircraft.
Integrating knowledge extracted from **text** (news, twits,...), **pictures**, **movies**...

- E.g., retrieve all documents and media where Bill Clinton makes statements about the Army

"Stripes," Dole, 1996

He (Bill Clinton) quickly became the first civilian commander-in-chief to salute his marine guards while entering or exiting an aircraft.
• Integrating knowledge extracted from **text** (news, twits,…), **pictures, movies**…

• E.g., retrieve all documents and media where Bill Clinton makes statements about the Army.

"Stripes," Dole, 1996

He (Bill Clinton) quickly became the first civilian commander-in-chief to salute his marine guards while entering or exiting an aircraft.

**what is a mention in a picture / movie?**

**how to represent it?**
### KnowledgeStore

**Beyond Text... and even more...**

- Interpreting / extracting / aligning knowledge from different media (e.g., video, commentary, text, ...)

<table>
<thead>
<tr>
<th>Frame</th>
<th>Commentary</th>
<th>Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="frame1.jpg" alt="Frame 1" /></td>
<td>“Sanchez, Sanchez,... goal. Sanchez equalizes for Chile”</td>
<td><code>dbpedia:Alexis_Sanchez</code> <code>scoreAt 32min</code></td>
</tr>
<tr>
<td><img src="frame2.jpg" alt="Frame 2" /></td>
<td>“Yellow card for the Chilean defender”</td>
<td><code>dbpedia:Mauricio_Pinilla</code> <code>yellowCardAt 102min</code></td>
</tr>
<tr>
<td><img src="frame3.jpg" alt="Frame 3" /></td>
<td>“Now is Marcelo turn, to kick the fourth penalty” “Marcelo... Goal”</td>
<td><code>dbpedia:Marcelo_Vieira</code> <code>kicks</code> <code>SuppPenalty4</code> <code>leadsTo goal</code></td>
</tr>
</tbody>
</table>
Thank You! Questions?

The 知識ストア Team

Roldano Cattoni, Francesco Corcoglioniti, Bernardo Magnini, Alessio Palmero Aprosio, Mohammed Qwaider, Marco Rospocher, Luciano Serafini

http://knowledgestore.fbk.eu