Semantic processing of text

Tools and applications

Peter Exner
Pierre Nugues

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.
Semantic Processing?

1. Sentence
   *Pierre wrote notes*

2. Logical representation
   `wrote(pierre, notes)`

3. Real world
   - Louis
   - Pierre
   - Charlotte

   refers to
   - operating systems
   - language processing
   - Prolog programming
Carsim

- A system to generate animated 3D scenes from a written text

```cpp
// Collision objects
ACCIDENT [
    COLLISION [
        ACTOR = vehicleB, front;
        VICTIM = vehicleA, leftside;
        COORD = (1.0, 1.0);
    ]
    COLLISION [
        ACTOR = vehicleA, front;
        VICTIM = tree1, unknown;
    ]
]
```
Carsim

<billkorpus.xml.4>
</billkorpus.xml.4>
## Semantic Parsing: CoNLL 2009

<table>
<thead>
<tr>
<th>Rank</th>
<th>Rank in task</th>
<th>System</th>
<th>Average</th>
<th>Catalan</th>
<th>Chinese</th>
<th>Czech</th>
<th>English</th>
<th>German</th>
<th>Japanese</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 (SRLonly)</td>
<td>Zhao</td>
<td>80.47</td>
<td>@ 80.32</td>
<td>77.72</td>
<td>85.19</td>
<td>85.44</td>
<td>75.99</td>
<td>78.15</td>
<td>@ 80.46</td>
</tr>
<tr>
<td>2</td>
<td>2 (SRLonly)</td>
<td>Nugues</td>
<td>80.31</td>
<td>80.01</td>
<td>@ 78.60</td>
<td>85.41</td>
<td>85.63</td>
<td>@ 79.71</td>
<td>76.30</td>
<td>76.52</td>
</tr>
<tr>
<td>3</td>
<td>1 (Joint)</td>
<td>Chen</td>
<td>79.96</td>
<td>80.10</td>
<td>76.77</td>
<td>82.04</td>
<td>@ 86.15</td>
<td>76.19</td>
<td>78.17</td>
<td>80.29</td>
</tr>
<tr>
<td>4</td>
<td>2 (Joint)</td>
<td>Che</td>
<td>79.94</td>
<td>77.10</td>
<td>77.15</td>
<td>@ 86.51</td>
<td>85.51</td>
<td>78.61</td>
<td>@ 78.26</td>
<td>76.47</td>
</tr>
<tr>
<td>5</td>
<td>3 (Joint)</td>
<td>Merlo</td>
<td>78.42</td>
<td>77.44</td>
<td>76.05</td>
<td>86.02</td>
<td>83.24</td>
<td>71.78</td>
<td>77.23</td>
<td>77.19</td>
</tr>
<tr>
<td>6</td>
<td>3 (SRLonly)</td>
<td>Meza-Ruiz</td>
<td>77.46</td>
<td>78.00</td>
<td>77.73</td>
<td>75.75</td>
<td>83.34</td>
<td>73.52</td>
<td>76.00</td>
<td>77.91</td>
</tr>
<tr>
<td>7</td>
<td>4 (Joint)</td>
<td>Bohnet</td>
<td>76.00</td>
<td>74.53</td>
<td>75.29</td>
<td>79.02</td>
<td>80.39</td>
<td>75.72</td>
<td>72.76</td>
<td>74.31</td>
</tr>
<tr>
<td>8</td>
<td>5 (Joint)</td>
<td>Asahara</td>
<td>75.65</td>
<td>72.35</td>
<td>74.17</td>
<td>84.69</td>
<td>84.26</td>
<td>63.66</td>
<td>77.93</td>
<td>72.50</td>
</tr>
<tr>
<td>9</td>
<td>6 (Joint)</td>
<td>Brown</td>
<td>72.85</td>
<td>72.18</td>
<td>72.43</td>
<td>78.02</td>
<td>80.43</td>
<td>73.40</td>
<td>61.57</td>
<td>71.95</td>
</tr>
<tr>
<td>10</td>
<td>7 (Joint)</td>
<td>Dai</td>
<td>70.78</td>
<td>66.34</td>
<td>71.57</td>
<td>75.50</td>
<td>78.93</td>
<td>67.43</td>
<td>71.02</td>
<td>64.64</td>
</tr>
<tr>
<td>11</td>
<td>8 (Joint)</td>
<td>Zhang</td>
<td>70.31</td>
<td>67.34</td>
<td>73.20</td>
<td>78.28</td>
<td>77.85</td>
<td>62.95</td>
<td>64.71</td>
<td>67.81</td>
</tr>
<tr>
<td>12</td>
<td>9 (Joint)</td>
<td>Lu Li</td>
<td>69.72</td>
<td>66.95</td>
<td>67.06</td>
<td>79.08</td>
<td>77.17</td>
<td>61.98</td>
<td>69.58</td>
<td>66.23</td>
</tr>
<tr>
<td>13</td>
<td>4 (SRLonly)</td>
<td>Baoli Li</td>
<td>69.26</td>
<td>74.06</td>
<td>70.37</td>
<td>57.46</td>
<td>69.63</td>
<td>67.76</td>
<td>72.03</td>
<td>73.54</td>
</tr>
<tr>
<td>14</td>
<td>10 (Joint)</td>
<td>Vallejo</td>
<td>68.95</td>
<td>70.14</td>
<td>66.71</td>
<td>71.49</td>
<td>75.97</td>
<td>61.01</td>
<td>68.82</td>
<td>68.48</td>
</tr>
<tr>
<td>15</td>
<td>5 (SRLonly)</td>
<td>Moreau</td>
<td>66.49</td>
<td>65.60</td>
<td>67.37</td>
<td>71.74</td>
<td>72.14</td>
<td>66.50</td>
<td>57.75</td>
<td>64.33</td>
</tr>
<tr>
<td>16</td>
<td>11 (Joint)</td>
<td>Luís</td>
<td>63.06</td>
<td>46.79</td>
<td>59.72</td>
<td>76.90</td>
<td>75.86</td>
<td>62.66</td>
<td>71.60</td>
<td>47.88</td>
</tr>
<tr>
<td>17</td>
<td>6 (SRLonly)</td>
<td>Täckström</td>
<td>61.27</td>
<td>57.11</td>
<td>63.41</td>
<td>71.05</td>
<td>67.64</td>
<td>53.42</td>
<td>54.74</td>
<td>61.51</td>
</tr>
<tr>
<td>18</td>
<td>7 (SRLonly)</td>
<td>Lin</td>
<td>57.18</td>
<td>61.70</td>
<td>70.33</td>
<td>60.43</td>
<td>65.66</td>
<td>59.51</td>
<td>23.78</td>
<td>58.87</td>
</tr>
<tr>
<td>19</td>
<td>12 (Joint)</td>
<td>Ren</td>
<td>56.69</td>
<td>41.00</td>
<td>72.58</td>
<td>62.82</td>
<td>67.56</td>
<td>54.31</td>
<td>58.73</td>
<td>39.80</td>
</tr>
<tr>
<td>20</td>
<td>13 (Joint)</td>
<td>Zeman</td>
<td>32.14</td>
<td>24.19</td>
<td>34.71</td>
<td>58.13</td>
<td>36.05</td>
<td>16.44</td>
<td>30.13</td>
<td>25.36</td>
</tr>
</tbody>
</table>

Semantic processing of text: Tools and applications / Peter Exner & Pierre Nugues
Coreference: CoNLL 2011

Official; Closed track; Predicted mentions

<table>
<thead>
<tr>
<th>System</th>
<th>MD F</th>
<th>MUC F&lt;sup&gt;1&lt;/sup&gt;</th>
<th>B-CUBED F&lt;sup&gt;2&lt;/sup&gt;</th>
<th>CEAFe&lt;sub&gt;m&lt;/sub&gt; F</th>
<th>CEAFe&lt;sub&gt;e&lt;/sub&gt; F&lt;sup&gt;3&lt;/sup&gt;</th>
<th>BLANC F</th>
<th>Official F&lt;sup&gt;1&lt;/sup&gt;+F&lt;sup&gt;2&lt;/sup&gt;+F&lt;sup&gt;3&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>lee</td>
<td>70.70</td>
<td>59.57</td>
<td>68.31</td>
<td>56.37</td>
<td>45.48</td>
<td>73.02</td>
<td>57.79</td>
</tr>
<tr>
<td>sapena</td>
<td>43.20</td>
<td>59.55</td>
<td>67.09</td>
<td>53.51</td>
<td>41.32</td>
<td>71.10</td>
<td>55.99</td>
</tr>
<tr>
<td>chang</td>
<td>64.28</td>
<td>57.15</td>
<td>68.79</td>
<td>54.40</td>
<td>41.94</td>
<td>73.71</td>
<td>55.96</td>
</tr>
<tr>
<td>nugues</td>
<td>68.96</td>
<td>58.61</td>
<td>65.46</td>
<td>51.45</td>
<td>39.52</td>
<td>71.11</td>
<td>54.53</td>
</tr>
<tr>
<td>santos</td>
<td>65.45</td>
<td>56.65</td>
<td>65.66</td>
<td>49.54</td>
<td>37.91</td>
<td>69.46</td>
<td>53.41</td>
</tr>
<tr>
<td>song</td>
<td>67.26</td>
<td>59.95</td>
<td>63.23</td>
<td>46.29</td>
<td>35.96</td>
<td>61.47</td>
<td>53.05</td>
</tr>
<tr>
<td>stoyanov</td>
<td>67.78</td>
<td>58.43</td>
<td>61.44</td>
<td>46.08</td>
<td>35.28</td>
<td>60.28</td>
<td>51.92</td>
</tr>
<tr>
<td>sobha</td>
<td>64.23</td>
<td>50.48</td>
<td>64.00</td>
<td>49.48</td>
<td>41.23</td>
<td>63.28</td>
<td>51.90</td>
</tr>
<tr>
<td>kobdani</td>
<td>61.03</td>
<td>53.49</td>
<td>65.25</td>
<td>42.70</td>
<td>33.79</td>
<td>62.61</td>
<td>51.04</td>
</tr>
<tr>
<td>zhou</td>
<td>62.31</td>
<td>48.96</td>
<td>64.07</td>
<td>47.53</td>
<td>39.74</td>
<td>64.72</td>
<td>50.92</td>
</tr>
<tr>
<td>charton</td>
<td>64.30</td>
<td>52.45</td>
<td>62.10</td>
<td>46.22</td>
<td>36.54</td>
<td>64.20</td>
<td>50.36</td>
</tr>
<tr>
<td>yang</td>
<td>63.93</td>
<td>52.31</td>
<td>62.32</td>
<td>46.55</td>
<td>35.33</td>
<td>64.63</td>
<td>49.99</td>
</tr>
<tr>
<td>hao</td>
<td>64.30</td>
<td>54.47</td>
<td>61.01</td>
<td>45.07</td>
<td>32.67</td>
<td>65.35</td>
<td>49.38</td>
</tr>
<tr>
<td>xinxin</td>
<td>61.92</td>
<td>46.62</td>
<td>61.93</td>
<td>44.75</td>
<td>36.23</td>
<td>64.27</td>
<td>48.46</td>
</tr>
<tr>
<td>zhang</td>
<td>61.13</td>
<td>47.28</td>
<td>61.14</td>
<td>44.46</td>
<td>35.19</td>
<td>65.21</td>
<td>48.07</td>
</tr>
<tr>
<td>kummerfeld</td>
<td>62.72</td>
<td>42.70</td>
<td>60.29</td>
<td>45.35</td>
<td>38.32</td>
<td>59.91</td>
<td>47.10</td>
</tr>
<tr>
<td>zhekovka</td>
<td>48.29</td>
<td>24.08</td>
<td>61.46</td>
<td>40.43</td>
<td>35.75</td>
<td>53.77</td>
<td>40.43</td>
</tr>
<tr>
<td>irwin</td>
<td>26.67</td>
<td>19.98</td>
<td>50.46</td>
<td>31.68</td>
<td>25.21</td>
<td>51.12</td>
<td>31.28</td>
</tr>
</tbody>
</table>
Constructing Large Proposition Databases

- Explore how semantic parsing can be scaled
- Parse a substantial part of Wikipedia
- Create large, semantically annotated, and multilingual proposition databases.
- Construct a querying interface
Inspiration: IBM Watson

source: http://www-03.ibm.com/innovation/us/watson/
Behind Watson: Automatic Learning

Automatic Learning From “Reading”

- Volumes of Text → Syntactic Frames → Semantic Frames
- Sentence Parsing → Generalization & Statistical Aggregation

- Inventors patent inventions (.8)
- Officials Submit Resignations (.7)
- People earn degrees at schools (.9)
- Fluid is a liquid (.6)
- Liquid is a fluid (.5)
- Vessels Sink (.7)
- People sink 8-balls (0.5) (in pool/0.8)

Complete Pipeline – English

Wiki Markup Filter → Sentence Detector → Tokenizer

Lemmatizer → POS Tagger → Dependency Parser

Semantic Role Labeler
"Ernest Miller Hemingway" (July 21, 1899 – July 2, 1961) was an American [[American literature|author]] and [[journalist]]. His distinctive writing style, characterized by economy and [[iceberg theory|understatement]], influenced 20th-century fiction, as did his life of adventure and public image. He produced most of his work between the mid-1920s and the mid-1950s. He won the [[Nobel Prize in Literature]] in 1954. Hemingway's fiction was successful because the characters he presented exhibited authenticity that resonated with his audience. Many of his works are classics of [[American literature]]. He published seven novels, six short story collections, and two non-fiction works during his lifetime; a further three novels, four collections of short stories, and three non-fiction works were published [[wikt:posthumous|posthumously]].

source: http://en.wikipedia.org/wiki/Ernest_Hemingway
Ernest Miller Hemingway (July 21, 1899 – July 2, 1961) was an American author and journalist. His distinctive writing style, characterized by economy and understatement, influenced 20th-century fiction, as did his life of adventure and public image. He produced most of his work between the mid-1920s and the mid-1950s. He won the Nobel Prize in Literature in 1954. Hemingway's fiction was successful because the characters he presented exhibited authenticity that resonated with his audience. Many of his works are classics of American literature. He published seven novels, six short story collections, and two non-fiction works during his lifetime; a further three novels, four collections of short stories, and three non-fiction works were published posthumously.
Ernest Miller Hemingway (July 21, 1899 – July 2, 1961) was an American author and journalist. His distinctive writing style, characterized by economy and understatement, influenced 20th-century fiction, as did his life of adventure and public image. He produced most of his work between the mid-1920s and the mid-1950s. **He won the Nobel Prize in Literature in 1954.** Hemingway's fiction was successful because the characters he presented exhibited authenticity that resonated with his audience. Many of his works are classics of American literature. He published seven novels, six short story collections, and two non-fiction works during his lifetime; a further three novels, four collections of short stories, and three non-fiction works were published posthumously.
Complete Pipeline – Tokenizer

He won the Nobel Prize in Literature in 1954.
He won the Nobel Prize in Literature in 1954.
Complete Pipeline – POS Tagger

He won the Nobel Prize in Literature in 1954.
He win the Nobel Prize in Literature in 1954.
He **won** the **Nobel Prize** in **Literature** in **1954**

**Predicate:** win.01

**Roles:**

- **Arg0:** winner
- **Arg1:** thing won (contest or prize)
- **Arg2:** beneficiary
- **Arg3:** loser, giver of prize
- **Arg4:** in-exchange-for
Athena

- A parallel parsing framework for parsing natural language.
  - Delegates parsing tasks to several computing nodes.
  - Uses a complete parsing pipeline, including a SRL.
Scaling up - challenges

- Network communication can be a bottleneck
- Data needs to be stored reliably
- Input needs to be evenly distributed
- Failed tasks need to be rerun
  - ...like running a huge fuzz testing case
Athena – How to parallelize?

- MPI
- Roll your own!
- Hadoop
What makes Hadoop useful?

- **Scalable**: Hadoop can reliably store and process petabytes.
- **Economical**: It distributes the data and processing across clusters of commonly available computers. These clusters can number into the thousands of nodes.
- **Efficient**: By distributing the data, Hadoop can process it in parallel on the nodes where the data is located. This makes it extremely rapid.
- **Reliable**: Hadoop automatically maintains multiple copies of data and automatically redeployes computing tasks based on failures.
Hadoop
Hadoop – Parsing Wikipedia

Article Selection Stage

Semantic Parsing stage

Seminar processing of text: Tools and applications / Peter Exner & Pierre Nugues
Crafoord cluster
Crafoord cluster

1 Master – JobTracker / NameNode
- Intel Xeon E5603 CPU
- 24GB ECC Memory
- ~12TB RAID 7 Storage

12 Workers – TaskTracker / DataNode
- Intel Core E5-1650 CPU (6-core /w HT)
- 32GB ECC Memory
- 240GB SSD Storage
Results – Parsing

- 3,835,000 articles (2012-02-11)
- 51,572,000 sentences
- 1,253,000,000 words (tokens)
- 198,000,000 propositions

- Effective Parsing Speed ~ 80,000 articles/day
Thermodynamic equations

Thermodynamics is expressed by a mathematical framework of *thermodynamic equations* which relate various thermodynamic quantities and physical properties measured in a laboratory or production process. Thermodynamics is based on an equation that describes the first law of thermodynamics.

<table>
<thead>
<tr>
<th>Argument 1</th>
<th>Predicate</th>
<th>Argument 2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>equation</td>
<td>describe.01</td>
<td>law</td>
<td>5</td>
</tr>
<tr>
<td>method</td>
<td>describe.01</td>
<td>approach</td>
<td>2</td>
</tr>
<tr>
<td>paper</td>
<td>describe.01</td>
<td>algorithm</td>
<td>1</td>
</tr>
</tbody>
</table>
Common-sense knowledge - high

- Males have income
- Schools include school
- Students attend schools
- Couple have children
- Municipality have population
- Municipality cover area
- Category contain articles
- Teams win championships
- Album sell copies
Common-sense knowledge - low

- Club finance improvements
- Doctors recover tissue
- Pelton design prototype
- McWilliam counterfeit cards
- Hurricane devastate islands
- Turrets have armor
- Weinberg recruit musicians
- Incident involve sinkholes
- Ricketts dominate season
Applications

• Resource for other NLP systems
• Information extraction
• Semantic search technology
• Machine translation
• Question & answer systems
• Provide statistics
Application: Search

Available at:

http://semantica.cs.lth.se/athena
Applications: Event Extraction

- Use semantic role labeling
- Use external resources
- Use Wikipedia as corpus
In 2001, Intel opened a plant in South Korea.

@prefix lode: <http://linkedevents.org/ontology/> .
...

natlang:Intel_opened
  a lode:Event ;
  lode:involvedAgent dbpedia:Intel ;
  lode:atTime [a time:DateTimeInterval ;
               time:hasDateTimeDescription
               [time:unitType time:unitYear ;
                time:year "2001"^^datatype:gYear ; ] ] ;
  lode:atPlace
    # geonames: 1835841 is South Korea
    <http://sws.geonames.org/1835841/> ;
  propbank:predicate "open.01" .
What is an event?

- Events need to have:
  - an **agent** with a DBpedia entry.
  - a **location** with a GeoNames or DBpedia entry.
  - a **date** with at least a starting year.

- 27,594 events extracted
  (from 10% of Wikipedia)
Applications: Entity Extraction

- Entities...
  - is a way of understanding text
  - are based on structured data
  - relate to each other
  - enable discovery and analysis
Google - Knowledge Graph

Source: http://www.google.com/insidesearch/features/search/knowledge.html
Google - Entity Search

Leonardo da Vinci - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Leonardo_da_Vinci
Leonardo di ser Piero da Vinci (April 15, 1452 – May 2, 1519, Old Style) was an Italian Renaissance polymath: painter, sculptor, architect, musician, ...
Personal life - Science and inventions - List of works - Mona Lisa

Leonardo da Vinci Biography - Facts, Birthday, Life Story ...
www.biography.com/people/leonardo-da-vinci-40396
Learn more about Leonardo da Vinci, the man who gave us the 'Mona Lisa' and 'The Last Supper,' on Biography.com.

Leonardo da Vinci - Museum of Science, Boston
legacy.mos.org/leonardo/
Provides a biography along with a multimedia section including images of his works.

European Commission - Leonardo da Vinci programme
europa.eu/education/lifelong-learning-programme/ldv_en.htm

Leonardo da Vinci
www.ucmp.berkeley.edu/history/vinci.html
Sometimes supernaturally, marvelously, they all congregate in one individual. . . . This was seen and acknowledged by all men in the case of Leonardo da Vinci, ...
Google – Entity Search

Which Rio?
Behind the Knowledge Graph

Leonardo da Vinci
From Wikipedia, the free encyclopedia

Leonardo di ser Piero da Vinci (Italian pronunciation: [leoˈnardo da ˈvintʃi]) (April 15, 1452 – May 2, 1519, Old Style) was an Italian Renaissance polymath: painter, sculptor, architect, musician, mathematician, engineer, inventor, anatomist, geologist, cartographer, botanist, and writer. His genius, perhaps more than that of any other figure, epitomized the Renaissance humanist ideal. Leonardo has often been described as the archetype of the Renaissance Man, as a person of "unquenchable curiosity" and "feverishly inventive imagination". He is widely considered to be one of the greatest painters of all time and perhaps the most diversely talented person ever to have lived. According to art historian Helen Gardner, the scope and depth of his interests were without precedent and "his mind and personality seem to us superhuman, the man himself mysterious and remote. Marco Rosci states that while there is much speculation about Leonardo, his vision of the world is essentially logical rather than mysterious, and that the empirical methods he employed were unusual for his time.

Leonardo da Vinci

<table>
<thead>
<tr>
<th>Name</th>
<th>Leonardo di ser Piero da Vinci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth</td>
<td>April 15, 1452</td>
</tr>
<tr>
<td>Nationality</td>
<td>Italian</td>
</tr>
<tr>
<td>Field</td>
<td>Many and diverse fields of the arts and sciences</td>
</tr>
<tr>
<td>Movement</td>
<td>High Renaissance</td>
</tr>
<tr>
<td>Works</td>
<td>Mona Lisa, The Last Supper, The Vitruvian Man, Lady with an Ermine</td>
</tr>
</tbody>
</table>

Semantic processing of text: Tools and applications / Peter Exner & Pierre Nugues
Approach

- Extend DBpedia
- Use Wikipedia as corpus

Semantic Role Labeling (LTH)

Coreference Resolution (Stanford CoreNLP)

Named Entity Linking (Wikifier)

Ontology Mapping (LTH)

Wikipedia
The Free Encyclopedia

DBpedia

DBpedia

- Dbpedia
  - A knowledgebase based on the structured information extracted from Wikipedia
  - An ontology describing almost 4 million things, covering 359 classes described by 1,775 properties.
  - Provides datasets in the shape of triples:

```
<dbpedia:Luc_Besson> <dbpedia-owl:birthDate> "1959-03-18"
```
Example – from text to DBpedia triple

Wikipedia:
“Luc Besson (born 18 March 1959) is a French film director, writer, and producer.”

DBpedia:
<dbpedia:Luc_Besson>  <dbpedia-owl:birthDate>  "1959-03-18"
Example step 1

“Luc Besson (born 18 March 1959) is a French film director, writer, and producer.”

“Besson was born in Paris to parents who were both Club Med scuba diving instructors.”

Semantic Role Labeling & Coreference Resolution

Besson was born in Paris to parents who were both Club Med scuba diving instructors.

Coreference chain

Semantic processing of text: Tools and applications / Peter Exner & Pierre Nugues
Example step 2

“Luc Besson (born 18 March 1959) is a French film director, writer, and producer. ”

“A1 bear.02 AM-TMP

“Besson was born in Paris to parents who were both Club Med scuba diving instructors.”

A1 bear.02 AM-LOC A2

Entity Extraction & Mention Entity Linking

URI from article
<dbpedia:Luc_Besson> <propbank:bear.02.AM-TMP> ”1959-03-18”

URI from Wikifier
<dbpedia:Besson> <propbank:bear.02.AM-LOC> <dbpedia:Paris>

Semantic processing of text: Tools and applications / Peter Exner & Pierre Nugues
Example step 3

URI from article
<dbpedia:Luc_Besson> <propbank:bear.02.AM-TMP> "1959-03-18"

URI from Wikifier
<dbpedia:Besson> <propbank:bear.02.AM-LOC> <dbpedia:Paris>

Coreference Inference & Ontology Mapping

<dbpedia:Luc_Besson> <dbpedia-owl:birthDate> "1959-03-18"

Coreference scoring module

<dbpedia:Luc_Besson> <dbpedia-owl:birthPlace> <dbpedia:Paris>
Ontology Mapping – learning step 1

“Besson married Milla Jovovich on 14 December 1997.”

Matching of Subject & Object with DBpedia Triplet

<dbpedia:Luc_Besson>  <propbank:marry.01.A1>  <dbpedia:Milla_Jovovich>

<dbpedia:Luc_Besson>  <dbpedia-owl:spouse>  <dbpedia:Milla_Jovovich>
Ontology Mapping – learning step 2

<dbpedia:Luc_Besson>  <dbpedia-owl:spouse>  <dbpedia:Milla_Jovovich>

Mapping

<dbpedia-owl:Person>  <propbank:marry.01.A1>  <dbpedia:Person>

<dbpedia-owl:spouse>
Ontology Mapping – reusing step 1

“On April 30, 2008, Carey married Cannon at her private estate…”

Triple extraction

\(<\text{dbpedia:Mariah_Carey}>\) \quad \langle\text{propbank:marry.01.A1}\rangle \quad \langle\text{dbpedia:Nick_Cannon}\rangle\)
Ontology Mapping – reusing step 2

“On April 30, 2008, Carey married Cannon at her private estate...”

Triple extraction

<dbpedia:Mariah_Carey>  <propbank:marry.01.A1>  <dbpedia:Nick_Cannon>

Subject & Object Generalization

<dbpedia-owl:Person>  <propbank:marry.01.A1>  <dbpedia-owl:Person>
"On April 30, 2008, Carey married Cannon at her private estate..."

Triple extraction

\[
\text{<dbpedia:Mariah_Carey>} \quad \text{<propbank:marry.01.A1>} \quad \text{<dbpedia:Nick_Cannon>}
\]

Subject & Object Generalization

\[
\text{<dbpedia-owl:Person>} \quad \text{<propbank:marry.01.A1>} \quad \text{<dbpedia-owl:Person>}
\]

Mapping

\[
\text{<dbpedia:Mariah_Carey>} \quad \text{<dbpedia-owl:spouse>} \quad \text{<dbpedia:Nick_Cannon>}
\]
Mappings learned

- In total 322 mappings

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>bear.02.AM-LOC</td>
<td>Place</td>
<td>birthPlace</td>
</tr>
<tr>
<td>Person</td>
<td>bear.02.AM-TMP</td>
<td>Date</td>
<td>birthDate</td>
</tr>
<tr>
<td>Person</td>
<td>marry.01.A1</td>
<td>Person</td>
<td>spouse</td>
</tr>
<tr>
<td>Place</td>
<td>locate.01.AM-LOC</td>
<td>Place</td>
<td>isPartOf</td>
</tr>
<tr>
<td>Place</td>
<td>open.01.AM-TMP</td>
<td>Date</td>
<td>openingDate</td>
</tr>
<tr>
<td>Place</td>
<td>build.01.AM-TMP</td>
<td>Numeric</td>
<td>yearOfConstruction</td>
</tr>
<tr>
<td>Organization</td>
<td>serve.01.A2</td>
<td>Place</td>
<td>broadcastArea</td>
</tr>
<tr>
<td>Organization</td>
<td>own.01.A1</td>
<td>Organization</td>
<td>subsidiary</td>
</tr>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>bandMember</td>
</tr>
</tbody>
</table>
Results - Triples

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total (from 114,895 articles)</td>
<td>1,023,316</td>
</tr>
<tr>
<td>DBpedia Mapped Triples</td>
<td>189,610</td>
</tr>
<tr>
<td>...of which 15,067 already exist in DBpedia</td>
<td>15,067</td>
</tr>
<tr>
<td>Unmapped Triples</td>
<td>833,706</td>
</tr>
</tbody>
</table>

- 833,706 triples couldn’t be mapped due to lacking a triple to learn from in Dbpedia, e.g.:

<dbpedia:ian_Holm> <propbank:portray.01> <dbpedia:Hercule_Poirot>

... interlinking with other datasets (LinkedMDB etc.) is needed!
Datasets available

An archive of entities in N-Triple format is available at:

http://semantica.cs.lth.se
Q&A
Polysemy

- How are polysemous mappings handled?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
<th>DBpedia</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>bandMember</td>
<td>20</td>
</tr>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>formerBandMember</td>
<td>11</td>
</tr>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>associatedMusicalArtist</td>
<td>8</td>
</tr>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>associatedBand</td>
<td>8</td>
</tr>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>keyPerson</td>
<td>6</td>
</tr>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>foundedBy</td>
<td>4</td>
</tr>
<tr>
<td>Organization</td>
<td>include.01.A2</td>
<td>Person</td>
<td>leader</td>
<td>1</td>
</tr>
</tbody>
</table>