Populating the Semantic Web with Historical Text

Kate Byrne, ICCS
Supervisors: Prof Ewan Klein, Dr Claire Grover

9th December 2008
Outline

Overview of My Research
- populating the Semantic Web
- the “Tether” System

Relation Extraction
- finding binary relations between NE pairs
- results

Mapping to RDF
- grounding relations in the wider graph
- dealing with generic nodes
Populating the Semantic Web

- Semantic Web born in 1994 – takeup slow
- Much of content is newly generated
- To add existing archives: need to expose as RDF
- Conversion to RDF not straightforward
Data from RCAHMS
The Royal Commission on the Ancient and Historical Monuments of Scotland

- Founded in February 1908
- [http://www.rcahms.gov.uk/](http://www.rcahms.gov.uk/)
- One of Scotland’s 6 National Collections
- The “memory keeper” for Scotland

**Mission** –
- **Survey** the built environment
- **Maintain a record** of buildings and archaeological sites
- **Promote understanding** of the material

Example: Informatics Forum, Aug 2007
Overview of Tether

Relational database

Published domain thesauri

Text documents
txt2rdf pipeline

Graph of triples

Relational Text documents of triples

Graph

Published domain thesauri

Graph of triples

Relational database

Published domain thesauri

Graph of triples

Relational database

Published domain thesauri

Graph of triples
Overview of Tether

Relational database

Published domain thesauri

Text documents

txt2rdf pipeline

Graph of triples

Relational Text documents of triples

Graph

Published domain thesauri

Relation Extraction

Mapping to RDF
NLP Work – *txt2rdf*

**Text documents**

- Pre-processing
  - Tokenise
  - Sentence and para split
  - POS tag

**Named Entity Recognition**

- Multi-word tokens and features
- List of NEs and classes
- Trained NER model

**Relation Extraction**

- Set of NE pairs and features
- List of relations and classes
- Trained RE model

**Graph of triples**

**RDF translation**

- Generate triples
- Remove unwanted relations
- Attach siteids

**Mapping to RDF**
NLP Work – *txt2rdf*

**Text documents**

- Tokenise
- Sentence and para split
- POS tag

**Pre-processing**

- Multi-word tokens and features
- List of NEs and classes
- Trained NER model

**Named Entity Recognition**

- Trained RE model
- Set of NE pairs and features
- List of relations and classes

**Relation Extraction**

- Generate triples
- Remove unwanted relations
- Attach site ids

**RDF translation**

- Graph of triples

**Mapping to RDF**

- Overview
- Relation Extraction
- Mapping to RDF
A Note on Evaluation

- Standard NLP metrics used – Precision, Recall, F-score:

\[
P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \quad F = \frac{2PR}{P + R}
\]

- **But**... precision actually preferred over recall
- End goal is Information Retrieval for non-experts
  ⇒ no information is better than false information
Overview of My Research
populating the Semantic Web
the “Tether” System

 Relation Extraction
finding binary relations between NE pairs
results

Mapping to RDF
grounding relations in the wider graph
dealing with generic nodes
Finding Binary Relations in Text

- NER as first step
- Special attention paid to NE nesting
- Then look for relations between pairs of NEs:
  - generate all possible pairings per document
  - add features
- Sequential tagger labels each pairing
Named Entity Recognition

- 11 categories:
  - ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT

- Unorthodox ones:
  - EVENT - verb phrases not noun phrases: visited, was found
  - SITETYPE, ARTEFACT, ROLE, EVENT – class terms

- Nesting:
  
  [[[Edinburgh]\text{PLACE}  University]\text{ORG}  Library]\text{ORG}
Named Entity Recognition

- 11 categories:
  - ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT
- Unorthodox ones:
  - EVENT - verb phrases not noun phrases: visited, was found
  - SITETYPE, ARTEFACT, ROLE, EVENT – class terms
- Nesting:
  - [[[Edinburgh]\text{PLACE} University]\text{ORG} Library]\text{ORG}
Named Entity Recognition

- 11 categories:
  - ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT

- Unorthodox ones:
  - EVENT - verb phrases not noun phrases: visited, was found
  - SITETYPE, ARTEFACT, ROLE, EVENT – class terms

- Nesting:
  [[[Edinburgh]^{PLACE} University]^{ORG} Library]^{ORG}
Named Entity Recognition

- 11 categories:
  - ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT
- Unorthodox ones:
  - EVENT - verb phrases not noun phrases: visited, was found
  - SITETYPE, ARTEFACT, ROLE, EVENT – class terms
- Nesting:
  \[
  [[[Edinburgh]^{PLACE} \text{ University}]^{ORG} \text{ Library}]^{ORG}
  \]
Relation Extraction

- Basic predicate categories:
  - eventRel, hasLocation, hasPeriod, instanceOf, partOf, sameAs, seeAlso
- $n$-ary eventRel predicate gets split up
- 11 binary predicates:
  - eventAgent, eventAgentRole, eventDate, eventPatient, eventPlace, hasLocation, hasPeriod, instanceOf, partOf, sameAs, seeAlso
RCAHMS Text with Relations Marked

[SOUTH WALLS], [MISBISTER], [[[THE {LOFTS}]]]

[ND38NW 29 centred 3325 8885]

Sites [[[recorded]]] during an [archaeological survey] undertaken on the lands of [[the {Loft}], [Longhope], as part of the pilot scheme for the [[[Historic {Scotland}]!]] {Farm} {Ancient} {Monument} Survey Grant Scheme].

Extracted Relations

• Examples of relations:
  • “The Loft” – hasLocation – Longhope
  • site – hasEvent – recording
  • recorded – hasLocation – “ND 3342 8884”
  • recorded – hasPatient – “Sub-rectangular cairn”

• RDF subject – property – object triples
## Results – NER Step

<table>
<thead>
<tr>
<th>Category</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADDRESS</td>
<td>82.40</td>
<td>81.61</td>
<td>82.00</td>
</tr>
<tr>
<td>ARTEFACT</td>
<td>75.83</td>
<td>18.06</td>
<td>29.17</td>
</tr>
<tr>
<td>DATE</td>
<td>95.12</td>
<td>82.08</td>
<td>88.12</td>
</tr>
<tr>
<td>EVENT</td>
<td>94.98</td>
<td>63.66</td>
<td>76.22</td>
</tr>
<tr>
<td>ORG</td>
<td>99.39</td>
<td>89.66</td>
<td>94.27</td>
</tr>
<tr>
<td>PERIOD</td>
<td>84.02</td>
<td>45.54</td>
<td>59.07</td>
</tr>
<tr>
<td>PERSNAME</td>
<td>96.71</td>
<td>74.82</td>
<td>84.37</td>
</tr>
<tr>
<td>PLACE</td>
<td>95.00</td>
<td>66.80</td>
<td>78.44</td>
</tr>
<tr>
<td>ROLE</td>
<td>98.00</td>
<td>54.44</td>
<td>70.00</td>
</tr>
<tr>
<td>SITENAME</td>
<td>64.55</td>
<td>61.20</td>
<td>62.83</td>
</tr>
<tr>
<td>SITETYPE</td>
<td>85.24</td>
<td>52.39</td>
<td>64.89</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>88.02</strong></td>
<td><strong>67.75</strong></td>
<td><strong>76.57</strong></td>
</tr>
</tbody>
</table>
Results – RE Step

<table>
<thead>
<tr>
<th>Relation</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
<th>Found</th>
</tr>
</thead>
<tbody>
<tr>
<td>eventAgent</td>
<td>98.42</td>
<td>98.70</td>
<td>98.56</td>
<td>3,794</td>
</tr>
<tr>
<td>eventAgentRole</td>
<td>69.23</td>
<td>30.00</td>
<td>41.86</td>
<td>13</td>
</tr>
<tr>
<td>eventDate</td>
<td>98.75</td>
<td>98.68</td>
<td>98.71</td>
<td>3,189</td>
</tr>
<tr>
<td>eventPatient</td>
<td>87.77</td>
<td>84.61</td>
<td>86.16</td>
<td>1,553</td>
</tr>
<tr>
<td>eventPlace</td>
<td>83.58</td>
<td>72.70</td>
<td>77.76</td>
<td>341</td>
</tr>
<tr>
<td>hasLocation</td>
<td>83.26</td>
<td>83.00</td>
<td>83.13</td>
<td>5,085</td>
</tr>
<tr>
<td>hasPeriod</td>
<td>83.69</td>
<td>73.86</td>
<td>78.47</td>
<td>233</td>
</tr>
<tr>
<td>instanceOf</td>
<td>52.00</td>
<td>31.52</td>
<td>39.25</td>
<td>100</td>
</tr>
<tr>
<td>partOf</td>
<td>78.87</td>
<td>51.38</td>
<td>62.22</td>
<td>568</td>
</tr>
<tr>
<td>sameAs</td>
<td>68.69</td>
<td>44.55</td>
<td>54.05</td>
<td>6,934</td>
</tr>
<tr>
<td>seeAlso</td>
<td>50.00</td>
<td>19.68</td>
<td>28.24</td>
<td>122</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>83.41</strong></td>
<td><strong>69.27</strong></td>
<td><strong>75.68</strong></td>
<td><strong>21,932</strong></td>
</tr>
</tbody>
</table>
RE Results for Event Relations

- EVENT category noted as unorthodox but...
- ...results are good
- Additional use for event extraction task:
  - populating RCAHMS relational database fields
Evaluating the Complete *txt2rdf* Pipeline

1. Use NE model to tag test set
2. Run RE model over “found” NE pairs
3. Evaluate against the gold standard
   - “new” relation pairs are FPs
   - every gold relation missed counts as FN
   - big variation across corpus: measure performance range
### Results for Full Pipeline

<table>
<thead>
<tr>
<th></th>
<th>“Hardest” data</th>
<th>“Easiest” data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>eventAgent</td>
<td>94.91</td>
<td>68.33</td>
</tr>
<tr>
<td>eventAgentRole</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>eventDate</td>
<td>80.69</td>
<td>57.19</td>
</tr>
<tr>
<td>eventPatient</td>
<td>83.33</td>
<td>4.00</td>
</tr>
<tr>
<td>eventPlace</td>
<td>36.36</td>
<td>8.33</td>
</tr>
<tr>
<td>hasLocation</td>
<td>67.90</td>
<td>59.31</td>
</tr>
<tr>
<td>hasPeriod</td>
<td>83.33</td>
<td>11.90</td>
</tr>
<tr>
<td>instanceOf</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>partOf</td>
<td>15.79</td>
<td>6.82</td>
</tr>
<tr>
<td>sameAs</td>
<td>47.63</td>
<td>16.92</td>
</tr>
<tr>
<td>seeAlso</td>
<td>18.18</td>
<td>13.64</td>
</tr>
<tr>
<td>Average</td>
<td>63.55</td>
<td>31.32</td>
</tr>
</tbody>
</table>

F-score range: 41.96% – 73.06%. Average: **57.51%**

Average precision: 73.35%
Overview of My Research
populating the Semantic Web
the “Tether” System

Relation Extraction
finding binary relations between NE pairs
results

Mapping to RDF
grounding relations in the wider graph
dealing with generic nodes
Mapping Text Relations to RDF

- “Bea Mill dates from the 19th century”
- “Bea Mill” – hasPeriod – “19th century”

@prefix : <http://www.ltg.ed.ac.uk/tether/> .
Grounding 1: Linking Text Relations to RCAHMS Sites

- “Bea Mill dates from the 19th century” [docid=3402]
Grounding 2: Connecting to Domain Thesauri

A Neolithic burial monument comprising a stone-built chamber within a mound of stones.

```
:Siteid#site123 :Classn/Sitetype#chambered+cairn
:hasClassn :skos:_prefLabel monThes:topTerm
:skos:scopeNoterdf:type
```

```
:Classn/Sitetype#monument+%28by+form%29
:Classn/Sitetype#religious+ritual+and+funerary
:Classn/Sitetype#chambered+tomb
:Classn/Sitetype#burial+cairn
:Classn/Sitetype#ring+cairn
:Classn/Sitetype#tomb
:Classn/Sitetype#square+cairn
```

```
"Passage Grave" "Heel Cairn" "Stalled Cairn"
```

```
:Classn/Sitetype#passage+grave
:Classn/Sitetype#stalled+cairn
:Classn/Sitetype#heel+cairn
```

```
:skos:broader
:skos:related
:skos:altLabel
:monThes:prefTerm
```
Grounding 2: Connecting to Domain Thesauri

"A Neolithic burial monument comprising a stone–built chamber within a mound of stones."

"CHAMBERED CAIRN"

"Heel Cairn"

"Stalled Cairn"

"Passage Grave"
Generic vs Specific Nodes

- Classes SITETYPE, ARTEFACT, ROLE, EVENT

- “Site 123 is a chambered cairn”

  ![Diagram]

  :Siteid#site123 :hasClassn :Classn/Sitetype#chambered+cairn

- “The chambered cairn is in Hoy and Graemsay”
Generic vs Specific Nodes

- Classes SITETYPE, ARTEFACT, ROLE, EVENT

- “Site 123 is a chambered cairn”

- “The chambered cairn is in Hoy and Graemsay”
Mapping Generic Categories

- “Site 101 was visited on 27 April 1969”
Subclass Labels

- **EVENT subclasses**: SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION, ALTERATION
- Annotated corpus includes NE subclass labels
  - was+found – `rdfs:subClassOf` – :Event/Find#
  - visited – `rdfs:subClassOf` – :Event/Visit#
  - built – `rdfs:subClassOf` – :Event/Creation#
- “Vocabulary” of EVENT subclasses available in graph
- Extracted text relations can be grounded in EVENT subclasses
Grounding 3: Placing EVENTs in Subclass Hierarchy

- :Siteid#site101
- :hasEvent
- :Event#visited101w117
- :hasPeriod
- :Time/Date#27+april+1969

- rdf:type :hasEvent
-rdf:subClassOf

- rdfs:subClassOf
- rdfs:label "Visited"

- "27 April 1969"
Summary

- 58% F-score for *txt2rdf* pipeline
- (Precision 73%)
- **Extracting structure from text is feasible**
- Class-valued categories:
  - members are class nodes when used generically
  - need unique ids when referring to specific context
- Using RDF makes integration easy –
  - with rest of site data
  - with domain thesauri
  - with further vocabularies in the future