Relation extraction and graph databases
— work in progress

27 July 2006
Outline

Revisions to SEER NE classes

Revisions to database schema

Candidate predicates

Integrating text relations with rest of graph
NER/NEC is preliminary step

- SEER-RCAHMS corpus — 1546 annotated files
- 9 NE classes, sub-divided into *Types* and *Subtypes*
- entity nesting: [[[Edinburgh] University] Library]
- summary of markup:
  - [http://homepages.inf.ed.ac.uk/s0233752/blog/refs/standoff.html](http://homepages.inf.ed.ac.uk/s0233752/blog/refs/standoff.html)
### NER/C results with SEER-RCAHMS corpus

<table>
<thead>
<tr>
<th>Entity Category</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINK</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ORG</td>
<td>97.12</td>
<td>95.30</td>
<td>96.20</td>
</tr>
<tr>
<td>PERIOD</td>
<td>78.60</td>
<td>67.14</td>
<td>72.42</td>
</tr>
<tr>
<td>PERSNAME</td>
<td>91.93</td>
<td>88.76</td>
<td>90.32</td>
</tr>
<tr>
<td>PLACE</td>
<td>79.13</td>
<td>87.88</td>
<td>83.27</td>
</tr>
<tr>
<td>REFERENCE</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>SITE</td>
<td>73.42</td>
<td>55.68</td>
<td>63.33</td>
</tr>
<tr>
<td>SIZE</td>
<td>92.64</td>
<td>93.37</td>
<td>93.00</td>
</tr>
<tr>
<td>TIMEX</td>
<td>93.72</td>
<td>95.21</td>
<td>94.46</td>
</tr>
<tr>
<td>Overall</td>
<td>86.18</td>
<td>83.83</td>
<td>84.99</td>
</tr>
</tbody>
</table>
Revisions

- Drop *Types* and *Subtypes*
- Drop SIZE class
- Split SITE - SITENAME, SITETYPE
- Expand some categories (ORG, CO-REF/LINK)...
- ... restrict some (DATE, REFERENCE)
- ... add new ones (ARTEFACT, ADDRESS)
- New list:
  - ORG, PERSNAME, SITETYPE, ARTEFACT, PLACE,
    SITENAME, ADDRESS, PERIOD, DATE, REFERENCE,
    CO-REF
- Revised annotation guidelines drafted
Automatic translation from RDBMS to RDF

<table>
<thead>
<tr>
<th>siteNo</th>
<th>name</th>
<th>parish</th>
<th>classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dirleton Castle</td>
<td>Dirleton</td>
<td>defence</td>
</tr>
<tr>
<td>2</td>
<td>Dirleton Cottage</td>
<td>Dirleton</td>
<td>residential</td>
</tr>
<tr>
<td>3</td>
<td>Drem Airfield</td>
<td>Dirleton</td>
<td>military</td>
</tr>
<tr>
<td>4</td>
<td>Jamie’s Neuk</td>
<td>Dirleton</td>
<td>military</td>
</tr>
</tbody>
</table>

- The devil is in the detail...
- See, eg, [Berners-Lee(2006)]
- Interest growing: http://esw.w3.org/topic/RdfAndSql
Moving complexity into the schema

- Group similar attributes/predicates together
  - eg location: \{parish, grid reference, street name,...\}
  - top level: loc, ind, id, sitename, date, classn, agent, desc
  - each with up to 4 sub-categories
- Now simplifying even further
Really Dreadful Framework

- What do we have on the right hand side?
- (site, location, Edinburgh) - *Edinburgh* has built-in semantics
- (site, rcmsgrade, AB3) - *AB3* is only meaningful in context
- Mixture of URIs and literals?
- What if a literal acquires “meaning” and needs its own “identity”?
- RDF tries to put semantics into data
Zooming in on the graph

- Everything’s either a literal or a URI
- Use bnodes to get round (some of) the URI ghastliness

We need this for every triple from the database, of the form:

```
http://clumsy.id/yuk#site
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/1999/02/22-rdf-syntax-ns#value
"AB3"
```

```
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/1999/02/22-rdf-syntax-ns#value
"AB3"
http://clumsy.if/fairly/transient#ind
```
Revised design

- Note *design* — no longer automatic
- The database fields become *classes* instead of *properties*
- Even *Edinburgh* becomes a literal hanging off a bnode
- Note also: we’re a long way from [Berners-Lee(2006)]

**Diagram:**

Instead of...

- `siteX`
- `rcmsgrade`
- `value` "AB3"

... or...

- `siteX`
- `rcmsgrade`
- `http://what.ever/durr#AB3`

... we have...

- `siteX`
- `ind`
- `value` "AB3"
- `type`
- `rcmsgrade`
Basic plan for determining predicate set

- Analyse SEER-RCAHMS corpus by POS tag — eg verb frequencies
- Cluster the candidates into groups — predicate classes
- Annotate for relations:
  - classification task
  - predicate recognition
  - predicate classification
- Attempt with heuristics; compare with machine learning
Analysing for candidate predicates

- Used chunker to find Verb Groups in SEER-RCAHMS corpus
- Listed by frequency:
  - [http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck3.freq](http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck3.freq)
- Converted to lower case and lemmatised:
  - [http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck2.freq](http://homepages.inf.ed.ac.uk/s0233752/blog/refs/vgCheck2.freq)
- 861 candidates
Clustering verbs

- Using NGD — Normalised Google Distance, [Cilibrasi and Vitanyi(2004)]
  \[ NGD(x, y) = \frac{\max\{\log f(x), \log f(y) - \log f(x, y)\}}{\log M - \min\{\log f(x), \log f(y)\}} \]

- Experiment with top 21 terms:
  file:///home/kate/phd/relPreds/ngdExp.html

- Using WordNet
  - find synset of candidate verb, and second hypernym
  - build inverted index: all candidates falling in second hypernym synset
  - perhaps should use antonyms?
  - comparison with NGD experiment:
    file:///home/kate/phd/relPreds/preds8.out

- by hand: file:///home/kate/phd/relPreds/vg21.byhand

- Possible alternative: pairwise comparison using wordnet::similarity
Tie NEs to database parent

- For RCAHMS, parent is *site* node
- Separate set of NE predicates
- Text relation → (NE, predicate, NE) triple

```
chambered cairn Henshall
describedBy
```

- NEorg
- NEsitetype
- NEpersname
- NEperiod
- siteXperiod
- id
- loc
- classn
- name
- period

**Diagram:**
- Site node with properties:
  - id
  - loc
  - classn
  - name
  - period
  - NEperiod
  - NEorg
  - NEsitetype
  - NEpersname

- NEs:
  - chambered cairn
  - Henshall

- DescribedBy relationship.
Nested entities

- Connect all NEs to parent site — flatten the nesting
- Use “partOf” to relate inner and outer entities

```
| date       | "1970" |
| person     | "Henshall" |
| type       | "Henshall, 1970" |
| partOf     | NEperson |
| value      | NEreference |
| partOf     | NEdate   |
| reference  | value    |
| type       | partOf   |
| value      | reference |
```

Diagram:
- person
- type
- "Henshall"
- value
- "Henshall, 1970"
- type
- partOf
- NEperson
- NEreference
- NEdate
- date
- type
- value
- "1970"
References

Tim Berners-Lee.
Relational Databases on the Semantic Web.
Internet note, 2006.
v 1.22 2006/02/01 (originally published September 1998).

Rudi Cilibrasi and Paul M. B. Vitanyi.
Automatic meaning discovery using Google.
Internet, 2004.