Nested Named Entity Recognition in Historical Archive Text

Kate Byrne, School of Informatics, University of Edinburgh

19 Sept 2007
Outline

Background
- Nature of the Data
- Problem and Proposed Solution

Text to RDF – NER Step
- Named Entity Recognition Setup
- Finding Nested Entities
- Results
Nature of the Data

- *The Royal Commission on the Ancient and Historical Monuments of Scotland*
- RCAHMS corpus: 1546 annotated texts on historical sites
- Entire RCAHMS dataset:
  - 250,000 records of archaeological and architectural sites
  - 1 text document per site
  - average of three archive items per site
  - “hybrid” data
- Cross domain testing with similar hybrid datasets (NLS, NMS)
Entity and Relation Annotation

- Corpus annotated to enable machine learning
- Annotation designed to be generic to cultural heritage:
  - Who? What? Where? When?
  - isA, seeAlso, sameAs, partOf
- Overall approach is generic across all hybrid datasets
  - annotation is the only domain-specific component
Entity and Relation Annotation

- Corpus annotated to enable machine learning
- Annotation designed to be generic to cultural heritage:
  - Who? What? Where? When?
  - isA, seeAlso, sameAs, partOf
- Overall approach is generic across all hybrid datasets
  - annotation is the only domain-specific component
A probable Neolithic house has been revealed by the removal of peat on the moor about 500 yards west of the bifurcation of the Vinsgarth-Stenness road, on the slightly sloping ground between the road and the Black Water. It consists of an oval setting of large boulders c. 40 m by 30 ft. overall, with the outer and inner faces of a wall, 7 1/2' thick traceable on the north and south arcs. The hollowed interior, generally associated with house-sites, is not seen but in the space there is a growth of peat. There is no sign of an entrance. Some 30 yds west there is a small oval enclosure which may be connected with the house. It is formed by only a single line of large stones set intermittently.

A Neolithic-Bronze Age homestead at HU 2296 7864 as described by Calder. There are traces of surrounding ruined field or enclosure walls, including the small enclosure mentioned by Calder. Probably an associated field system. Homestead and enclosure surveyed at 1/2500. Visited by OS (NKB) 23 May 1969.

Data Access Problems

Objective:

Open up cultural heritage data for the general user

Problems:

1. Complex database structure
2. Specialist terminology
3. Limited access to text content
The Proposal

- Transform hybrid data into directed graph: **datagraph**

  ![Graph Diagram]

  - Components:
    - structured database fields
    - domain thesauri
    - text documents

  - Datagraph characteristics:
    - graph of binary relations, expressed as RDF triples
    - nodes are **NEs**, thesaurus terms, database field contents
    - edges are relation predicates, database field names
    - not necessarily consistent across entire extent
The Proposal

• Transform hybrid data into directed graph: *datagraph*

![Diagram of a directed graph with subject, predicate, and object nodes.]

• Components:
  - structured database fields
  - domain thesauri
  - text documents

• Datagraph characteristics:
  - graph of binary relations, expressed as RDF triples
  - nodes are NEs, thesaurus terms, database field contents
  - edges are relation predicates, database field names
  - not necessarily consistent across entire extent
The Proposal

- Transform hybrid data into directed graph: *datagraph*

- Components:
  - structured database fields
  - domain thesauri
  - text documents

- Datagraph characteristics:
  - graph of binary relations, expressed as RDF triples
  - nodes are *NEs*, thesaurus terms, database field contents
  - edges are *relation predicates*, database field names
  - not necessarily consistent across entire extent
Claims

1. Deals with the three problems (structure, terminology, text)
2. **Text can be adequately realised as graph of binary relations**
3. Extra gains:
   - use graph locality to deal with inconsistency
   - potential discovery of latent relationships
   - graph summaries of intermediate results...
   - ...enabling guided queries
Text to RDF Conversion

Step 1. NER – identify and classify node terms
Step 2. RE – find relations between these
Named Entity Recognition

- NE classes:
  - org, persname, role, sitetype, artefact, sitename, place, address, period, date, event

- event subclasses:
  - survey, excavation, find, visit, description, creation, alteration

- Using supervised tagging (following pre-processing)
- C&C classifier, tuned for NER [Curran and Clark, 2003]
Background

Text to RDF – NER Step

Named Entity Recognition

- **NE classes:**
  
  org, persname, role, sitetype, artefact, sitename, place, address, period, date, event

- **event subclasses:**
  
  survey, excavation, find, visit, description, creation, alteration

- Using supervised tagging (following pre-processing)
- C&C classifier, tuned for NER [Curran and Clark, 2003]
Named Entity Recognition

- **NE classes:**
  - org, persname, role, sitetype, artefact, sitename, place, address, period, date, event

- **Event subclasses:**
  - survey, excavation, find, visit, description, creation, alteration

- Using supervised tagging (following pre-processing)
- C&C classifier, tuned for NER [Curran and Clark, 2003]
Nested Entities

- RCAHMS corpus: 10% of NEs have others nested within them
- Up to three levels of nesting in corpus, e.g.

```
[[[Edinburgh]^{PLACE} University]^{ORG} Library]^{ORG} is adjacent to
[[Adam Ferguson]^{PERSNAME} Building]^{ADDRESS}
```

- If nested NEs not found, relations involving them are lost
  - `hasLocation(Adam Ferguson Building, Edinburgh)`
- If they are found, “intra-relations” come (more or less) free
  - `partOf(Edinburgh University Library, Edinburgh University)`
  - `hasLocation(Edinburgh University, Edinburgh)`
Nested Entities

- RCAHMS corpus: 10% of NEs have others nested within them
- Up to three levels of nesting in corpus, e.g.

  [[[Edinburgh]\textsuperscript{PLACE} University]\textsuperscript{ORG} Library]\textsuperscript{ORG} is adjacent to
  [[Adam Ferguson]\textsuperscript{PERSNAME} Building]\textsuperscript{ADDRESS}

- If nested NEs not found, relations involving them are lost
  - hasLocation(Adam Ferguson Building, Edinburgh)

- If they are found, “intra-relations” come (more or less) free
  - partOf(Edinburgh University Library, Edinburgh University)
  - hasLocation(Edinburgh University, Edinburgh)
Nested Entities

• RCAHMS corpus: 10% of NEs have others nested within them
• Up to three levels of nesting in corpus, e.g.

  [[[Edinburgh]]^{PLACE} University]^{ORG} Library]^{ORG} is adjacent to
  [[Adam Ferguson]]^{PERSNAME} Building]^{ADDRESS}

• If nested NEs not found, relations involving them are lost
  • hasLocation(Adam Ferguson Building, Edinburgh)
• If they are found, “intra-relations” come (more or less) free
  • partOf(Edinburgh University Library, Edinburgh University)
  • hasLocation(Edinburgh University, Edinburgh)
Standard Tagging

- Standard classifier: one label per token
- Only one layer of NEs can be found:

as O
Edinburgh B-ORG
University I-ORG
Library I-ORG
is O
adjacent O
to O
Adam B-ADDRESS
Ferguson I-ADDRESS
Building I-ADDRESS
Finding Nested Entities

- Possible approaches:
  - cascading and layering, combining results
  - multilabel tagging [McDonald et al., 2005]
  - joined label tagging plus cascading [Alex et al., 2007]

- My approach: multi-word tokenisation
  - combine tokens so that one label per token works
- Compare against standard single token tagging
Finding Nested Entities

- Possible approaches:
  - cascading and layering, combining results
  - multilabel tagging [McDonald et al., 2005]
  - joined label tagging plus cascading [Alex et al., 2007]
- My approach: multi-word tokenisation
  - combine tokens so that one label per token works
- Compare against standard single token tagging
Finding Nested Entities

- Possible approaches:
  - cascading and layering, combining results
  - multilabel tagging [McDonald et al., 2005]
  - joined label tagging plus cascading [Alex et al., 2007]

- My approach: multi-word tokenisation
  - combine tokens so that one label per token works

- Compare against standard single token tagging
Multi-word Tokenisation

as O
as_Edinburgh O
as_Edinburgh_University O
Edinburgh PLACE
Edinburgh_University ORG
Edinburgh_University_Library ORG
University O
University_Library O
University_Library_is O
Library O
Library_is O
Library_is_adjacent O
NER Results

<table>
<thead>
<tr>
<th>Using C&amp;C</th>
<th>P %</th>
<th>R %</th>
<th>F %</th>
<th>Correct NEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best single-token run</td>
<td>76.98</td>
<td>75.18</td>
<td>76.07</td>
<td>18,379</td>
</tr>
<tr>
<td>Multi-token, unweighted</td>
<td>87.70</td>
<td>66.79</td>
<td>75.83</td>
<td>18,322</td>
</tr>
<tr>
<td>Multi-token, smoothed</td>
<td>78.43</td>
<td>75.91</td>
<td>77.15</td>
<td>20,825</td>
</tr>
</tbody>
</table>

- Performance comparable to method that only finds single level
- Precision improved at expense of recall – good!
- Smoothed model outputs 13% extra NEs
- Main drawback: much slower
Results Using Untuned Classifier

<table>
<thead>
<tr>
<th>Using ZLMaxent</th>
<th>P %</th>
<th>R %</th>
<th>F %</th>
</tr>
</thead>
<tbody>
<tr>
<td>single-word tokens</td>
<td>41.06</td>
<td>48.56</td>
<td>44.49</td>
</tr>
<tr>
<td>multi-word tokens</td>
<td>78.59</td>
<td>46.90</td>
<td>58.75</td>
</tr>
</tbody>
</table>

- Simple experiment using classifier not optimised for NER
- Performance significantly higher with multi-tokens
- Bias towards precision confirmed
- Very fast
Summary

- NER is only one component of pipeline
- Goal is to capture text content as RDF triples
- Important to deal with nested entities
  - “free” relations wherever nesting occurs
- Multi-word tokenisation:
  - conceptually simple
  - easy to automate
  - results comparable to more sophisticated approaches
  - 13% extra NEs found
References

