Automatic Extraction of Archaeological Events from Text

Kate Byrne and Ewan Klein, University of Edinburgh

25th March 2009

Outline

Overview

– project setting and data used

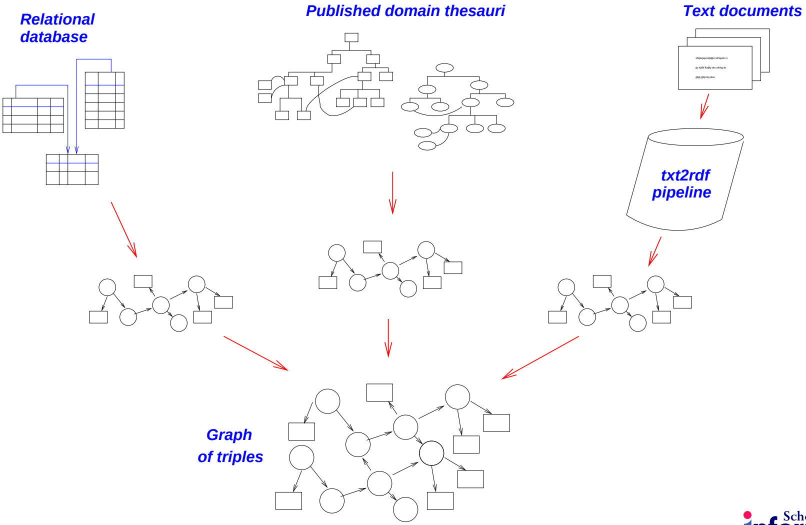
Event Extraction

– finding relations between textual entity mentions

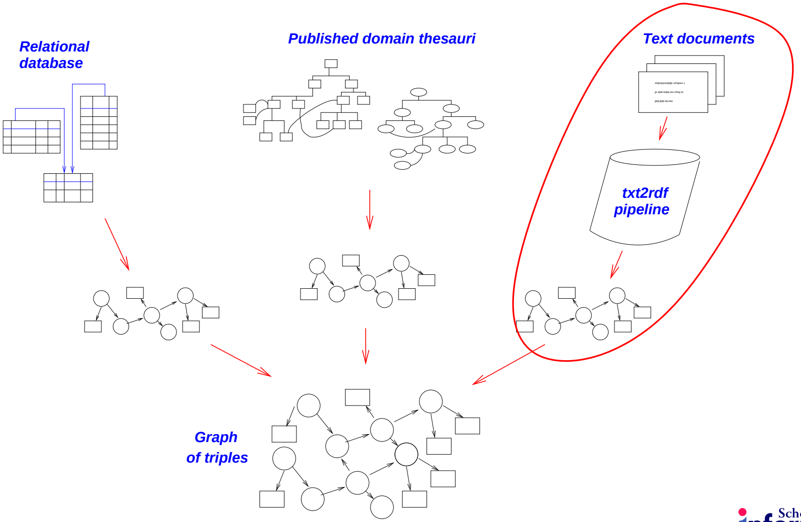
Results

– for NER, RE and the combination

Overview of *Tether*



Overview of *Tether*



Data from RCAHMS

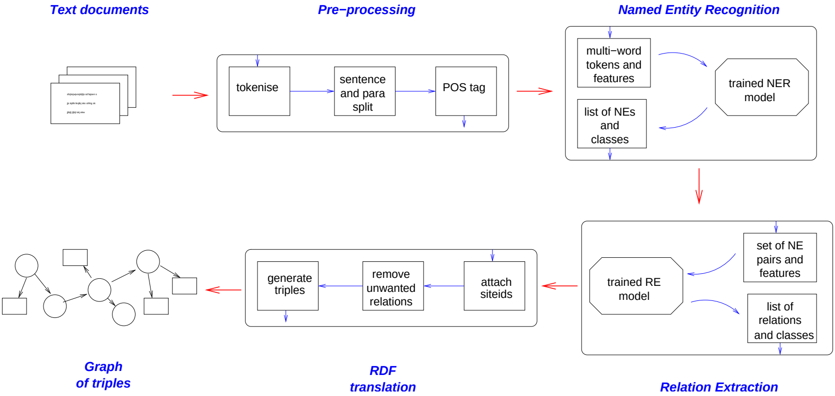
- The "memory keeper" for Scotland
- http://www.rcahms.gov.uk/
- One of Scotland's 6 National Collections

Royal Commission on the Ancient and Historical Monuments of Scotland

- Recording Scotland's places, from the Neolithic to Now:
  - Skara Brae
  - Informatics Forum

- Extracting events is a requirement

NLP Work – *txt2rdf*



Finding Binary Relations in Text

- Named Entity Recognition 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 –
    NE classes, word separation, POS tags, nesting, in sentence...
- Sequential tagger labels each pairing

Supervised Learning for Relation Extraction

The following were found in Unst by Mr A T Cluness : a steatite dish , ...

Supervised Learning for Relation Extraction

FIND EVENT PLACE PERSNAME ARTEFACT

The following were found in Unst by Mr A T Cluness : a steatite dish , ...

Supervised Learning for Relation Extraction

FIND EVENT PLACE PERSNAME ARTEFACT

The following were found in Unst by Mr A T Cluness : a steatite dish , ...

were_found	unst
were_found	a_t_cluness
were_found	steatite_dish
unst	a_t_cluness
unst	steatite_dish
a_t_cluness	steatite_dish

Supervised Learning for Relation Extraction

FIND EVENT *PLACE* *PERSNAME* *ARTEFACT*

The following were found in Unst by Mr A T Cluness : a steatite dish , ...

eventLocation	were_found	unst
eventAgent	were_found	a_t_cluness
eventPatient	were_found	steatite_dish
O	unst	a_t_cluness
O	unst	steatite_dish
O	a_t_cluness	steatite_dish

Supervised Learning for Relation Extraction

FIND EVENT PLACE PERSNAME ARTEFACT

The following were found in Unst by Mr A T Cluness : a steatite dish , ...

eventLocation	were_found	unst	*cls1=event*	*cls2=place*	*wdsep=+2 ...*
eventAgent	were_found	a_t_cluness	*cls1=event*	*cls2=persname*	*wdsep=+5 ...*
eventPatient	were_found	steatite_dish	*cls1=event*	*cls2=artefact*	*wdsep=+9 ...*
O	unst	a_t_cluness	*cls1=place*	*cls2=persname*	*wdsep=+9 ...*
O	unst	steatite_dish	*cls1=place*	*cls2=artefact*	*wdsep=+9 ...*
O	a_t_cluness	steatite_dish	*cls1=persname*	*cls2=artefact*	*wdsep=+9 ...*

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]$^{\mathrm{PLACE}}$ University]$^{\mathrm{ORG}}$ Library]$^{\mathrm{ORG}}$

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}

School of informatics

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}

Relation Extraction

- Basic predicate categories:

  eventRel, hasLocation, hasPeriod, instanceOf, partOf, sameAs, seeAlso

- *n*-ary eventRel predicate:

  eventAgent, eventAgentRole, eventDate, eventPatient, eventPlace

- event types: SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION, ALTERATION

Relation Extraction

- Basic predicate categories:

  eventRel, hasLocation, hasPeriod, instanceOf, partOf, sameAs, seeAlso

- *n*-ary eventRel predicate:

  eventAgent, eventAgentRole, eventDate, eventPatient, eventPlace

- event types: SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION, ALTERATION

Relation Extraction

- Basic predicate categories:

  eventRel, hasLocation, hasPeriod, instanceOf, partOf, sameAs, seeAlso
- *n*-ary eventRel predicate:

  eventAgent, eventAgentRole, eventDate, eventPatient, eventPlace
- event types: SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION, ALTERATION

Mapping Text Relations to RDF

site456

{[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] . [ND 3311 8890] Two [small {cairns]} . [ND 3336 8889] {[Cairn]} . [ND 3339 8885] {[Cairn]} . [ND 3339 8886] {[Clearance cairn]} . [ND 3342 8884] [Sub-rectangular cairn]} . [ND 3339 8883] {[Well]} Sponsors : {[Historic {[Scotland]]}} , {[M J Jones]} . {{[[N Card]} {[1998]]

Mapping Text Relations to RDF

site456

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

[ND38NW 29 centred 3325 8885]

event

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] . [ND 3311 8890] Two [small {cairns]} . [ND 3336 8889] {[Cairn]} . [ND 3339 8885] {[Cairn]} . [ND 3339 8886] {[Clearance cairn]} . [ND 3342 8884] [Sub-rectangular cairn]} . [ND 3339 8883] {[Well]} Sponsors : {[Historic {[Scotland]]}} , {[M J Jones]} . {{[[N Card]} {[1998]]

eventPatient

eventPlace

Mapping Text Relations to RDF

site456

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

[ND38NW 29 centred 3325 8885]

event

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] . [ND 3311 8890] Two [small {cairns]} . [ND 3336 8889] {[Cairn]} . [ND 3339 8885] {[Cairn]} . [ND 3339 8886] {[Clearance cairn]} . [ND 3342 8884] [Sub-rectangular cairn]} . [ND 3339 8883] {[Well]} Sponsors : {[Historic {[Scotland]]}} , {[M J Jones]} . {{[[N Card]} {[1998]]

eventPatient

eventPlace

site456 − hasEvent − recordingX
recordingX − hasLocation − "ND 3342 8884"
recordingX − hasPatient − "Sub−rectangular cairn"

Results – NER Step

	Precision %	Recall %	F-score %	Count
ADDRESS	82.40	81.61	82.00	3,458
PLACE	95.00	66.80	78.44	2,503
SITENAME	64.55	61.20	62.83	2,712
DATE	95.12	82.08	88.12	3,519
PERIOD	84.02	45.54	59.07	400
EVENT	**94.98**	**63.66**	**76.22**	**3,176**
ORG	99.39	89.66	94.27	2,730
PERSNAME	96.71	74.82	84.37	2,318
ROLE	98.00	54.44	70.00	90
SITETYPE	85.24	52.39	64.89	5,668
ARTEFACT	75.83	18.06	29.17	879
Average	88.02	67.75	**76.57**	(27,453)

IAA F-score 78.09%

Results – Extracting Event Relations

Relation	Precision %	Recall %	F-score %	Found
eventAgent	98.42	98.70	98.56	3,794
eventAgentRole	69.23	30.00	41.86	13
eventDate	98.75	98.68	98.71	3,189
eventPatient	87.77	84.61	86.16	1,553
eventPlace	83.58	72.70	77.76	341
Events Average	87.55	76.94	**80.61**	(8,890)
Overall Average	83.41	69.27	**75.68**	(21,932)

- IAA F-score 82.51%
- Note deliberate preference of Precision over Recall

Results for Full Pipeline (Event Relations)

Relation	Avg Precision	Avg Recall	Avg F-score
eventAgent	97.46	82.18	88.72
eventAgentRole	0.00	0.00	0.00
eventDate	87.75	71.73	78.64
eventPatient	90.69	42.99	48.46
eventPlace	36.36	17.33	27.62
Overall Average	73.35	48.24	57.51

- Evaluation over only 10% of corpus, so sparse categories lost
- eventAgent and eventDate are largest categories

Summary

- Event modelling is unorthodox in NER terms but results good
  - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
  - average 81% F-score for event relations (overall avg: 72%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible

Summary

- Event modelling is unorthodox in NER terms but results good
  - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
  - average 81% F-score for event relations (overall avg: 76%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible

Summary

- Event modelling is unorthodox in NER terms but results good
  - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
  - average 81% F-score for event relations (overall avg: 76%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible

School of informatics

Summary

- Event modelling is unorthodox in NER terms but results good
  - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
  - average 81% F-score for event relations (overall avg: 76%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible

Summary

- Event modelling is unorthodox in NER terms but results good
  - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
  - average 81% F-score for event relations (overall avg: 76%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible

Summary

- Event modelling is unorthodox in NER terms but results good
  - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
  - average 81% F-score for event relations (overall avg: 76%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible

Summary

- Event modelling is unorthodox in NER terms but results good
  - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
  - average 81% F-score for event relations (overall avg: 76%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible

Summary

- Event modelling is unorthodox in NER terms but results good
    - EVENT NE recognition: 76% F-score (avg: 77%)
- Event relations are **easier** than others:
    - average 81% F-score for event relations (overall avg: 76%)
- Models deliberately trained to favour Precision over Recall
- Extraction to RDF graph, as shown...
- ...or to populate RDB tables if desired
- Automatic extraction of events from text is feasible