Automatic Extraction of Archaeological Events from Text

Kate Byrne Ewan Klein

University of Edinburgh

Presented by
Mainack Mondal

Course supervisor
Dr. Caroline Sporleder

How to represent archeological data?

RCAHMS
Memory keeper for Scotland

Royal Commission on the Ancient and Historical Monuments of Scotland

Skara Brae
(3180 BCE)

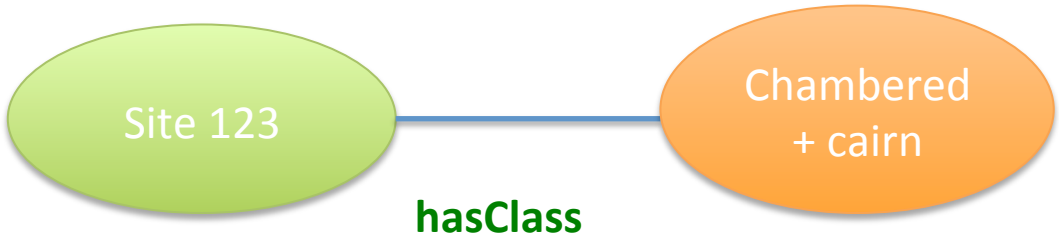
Faculy building
(21th Century)

Question: “Skara Brae **was found** at _____”

Automated extraction of events is a requirement

Automatic extraction of events

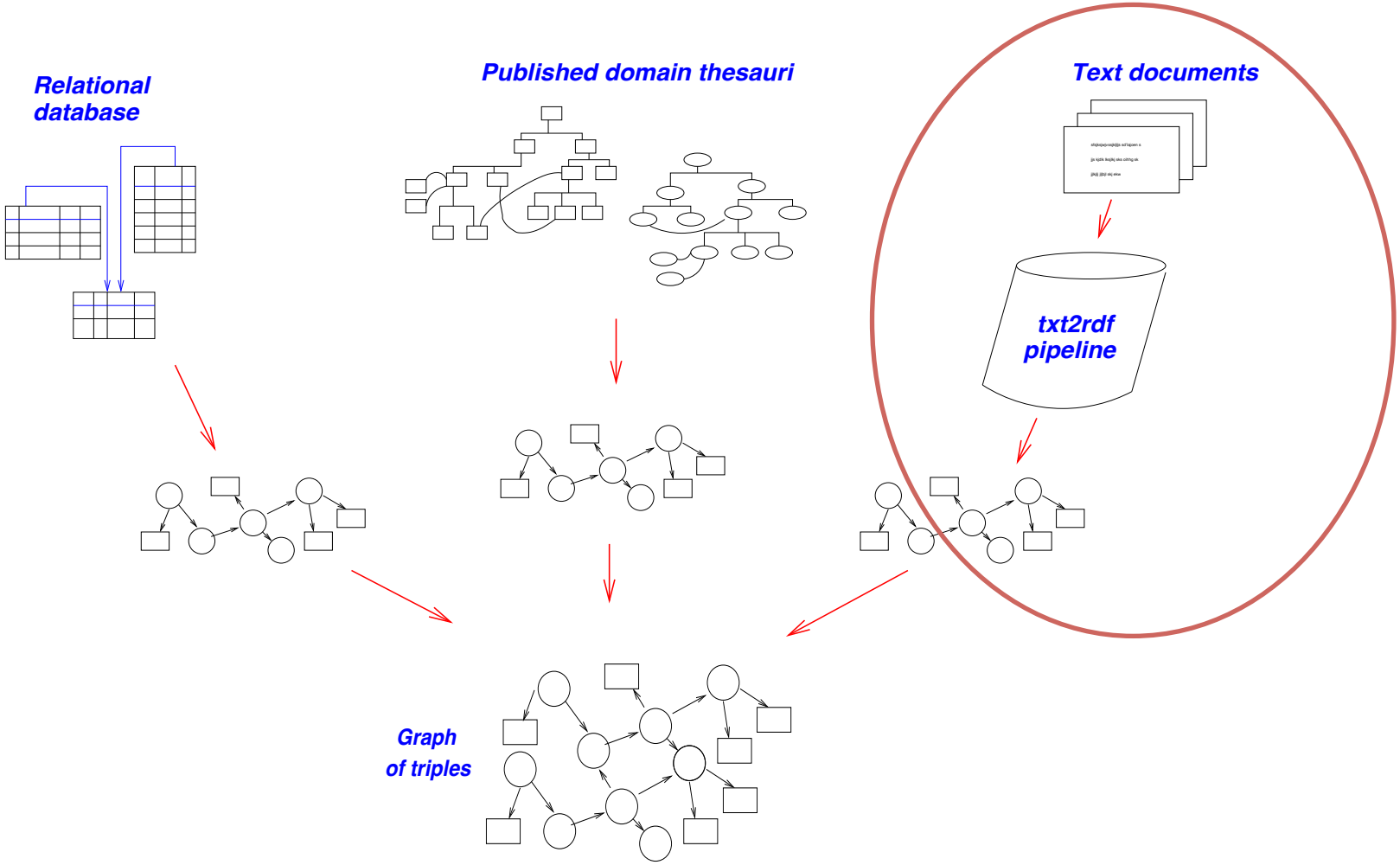
- Idea: Semantic web is useful
  - **site123** is **classified as** a **chambered cairn**



- Resource Description Format (RDF)
  - Subject – Predicate – Object

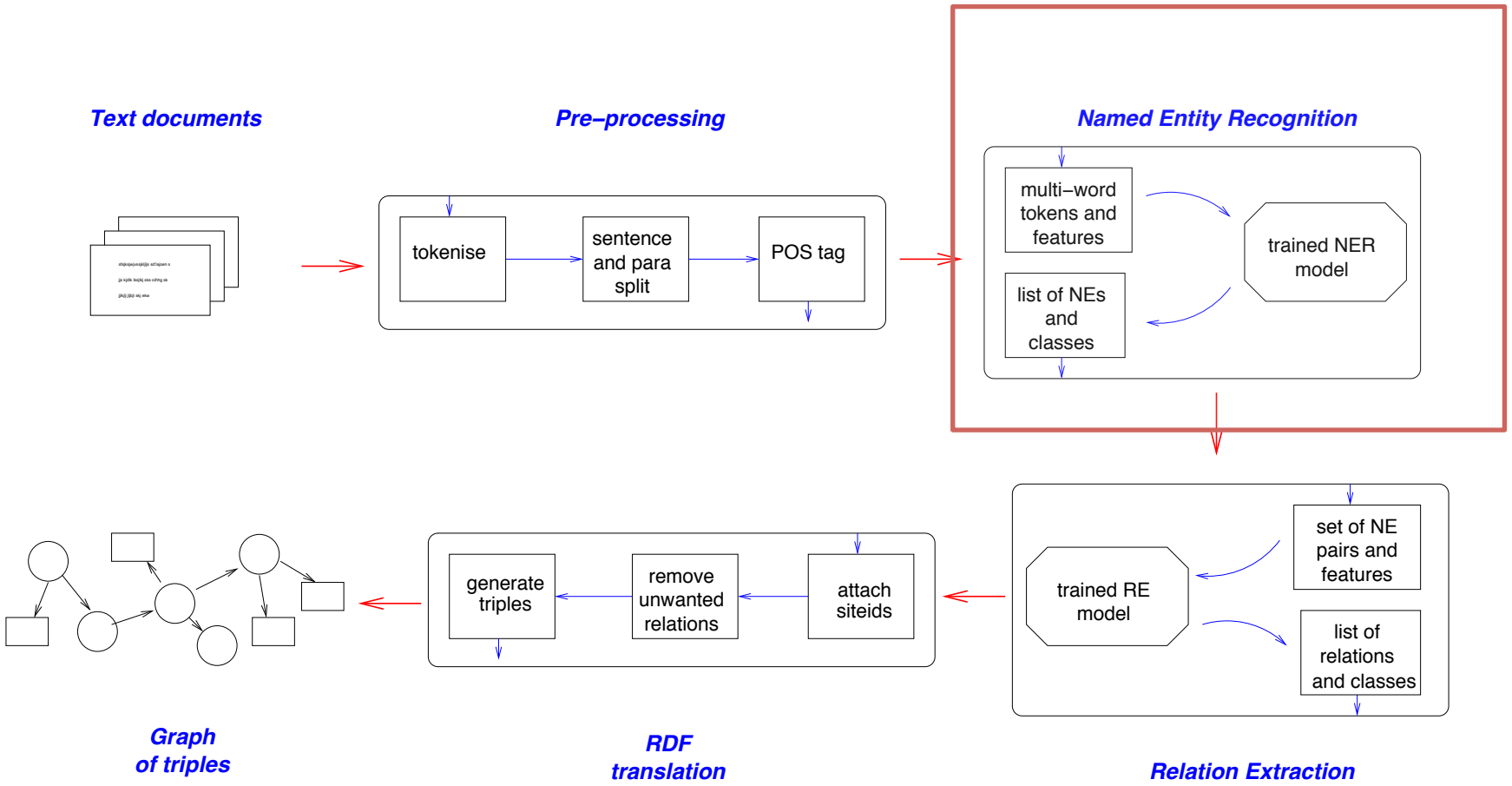
- How to convert text data to RDF format

Tether: converting RCAHMS data to RDF format



Img source: authors' presentation

txt2rdf: the pipeline



Img source: authors' presentation

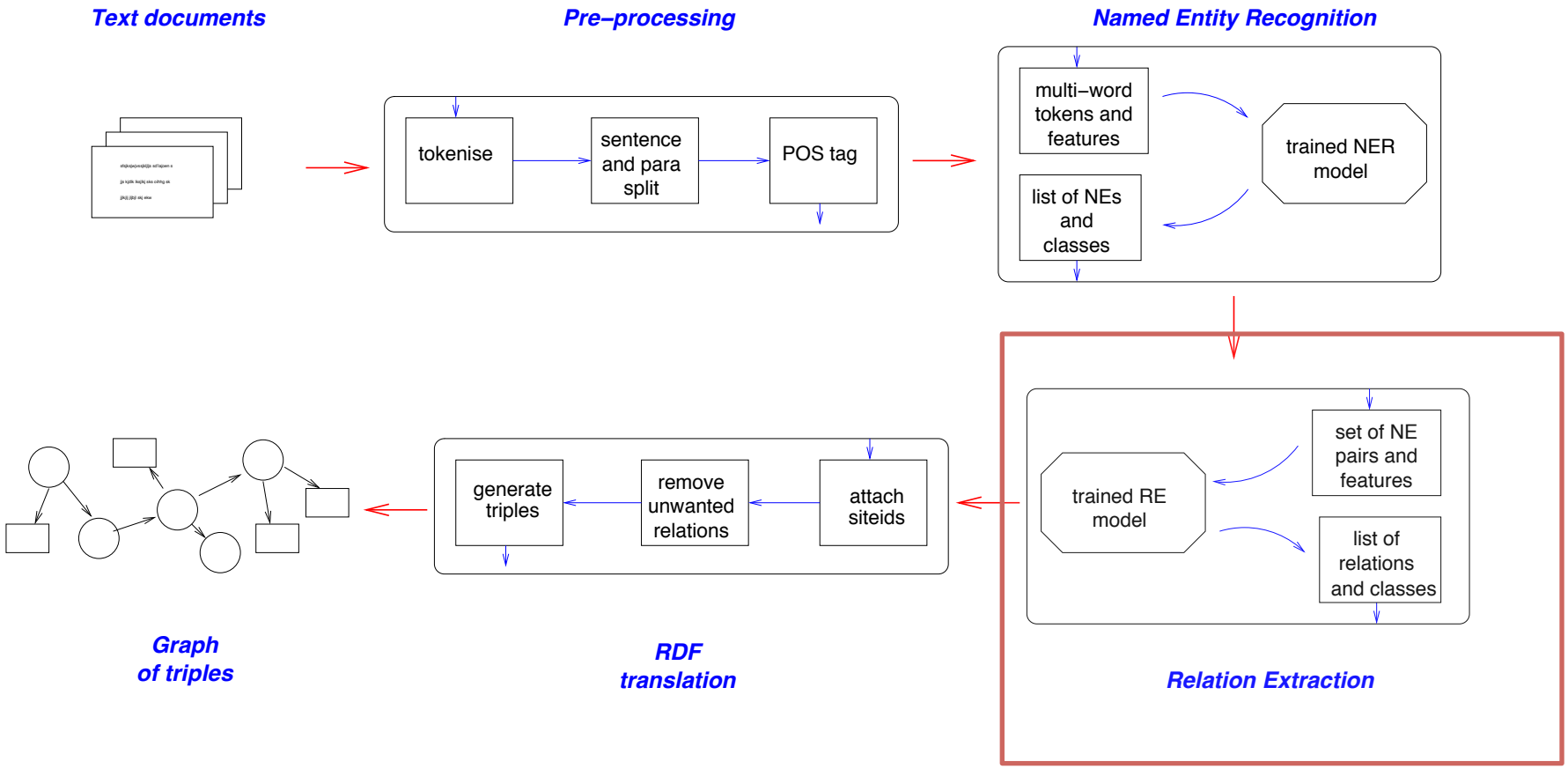
Named Entity recognition

List of Categories (ORG, COUNTRY…)

[DE] COUNTRY
[UdS] ORG

- 11 categories:
  - ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT
- Unorthodox ones:
  - EVENT – SURVEY, EXCAVATION, FIND
- Nesting:
  - [[[Edinburgh]^{PLACE} University]^{ORG} Library]^{ORG}

txt2rdf

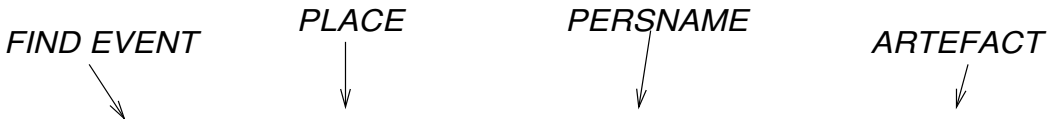


Img source: authors' presentation

Finding binary relations in text

- Named Entity Recognition was 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...

Supervised learning for Relation extraction



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...*

Img source: authors' presentation

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

Working Example for txt2rdf

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]]

event

eventPlace

eventPatient

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

Img source: authors' presentation

Results: evaluating 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)

table source: authors' presentation

Results: evaluating RE step

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)

table source: authors presentation

Results: evaluating full txt2rdf pipeline

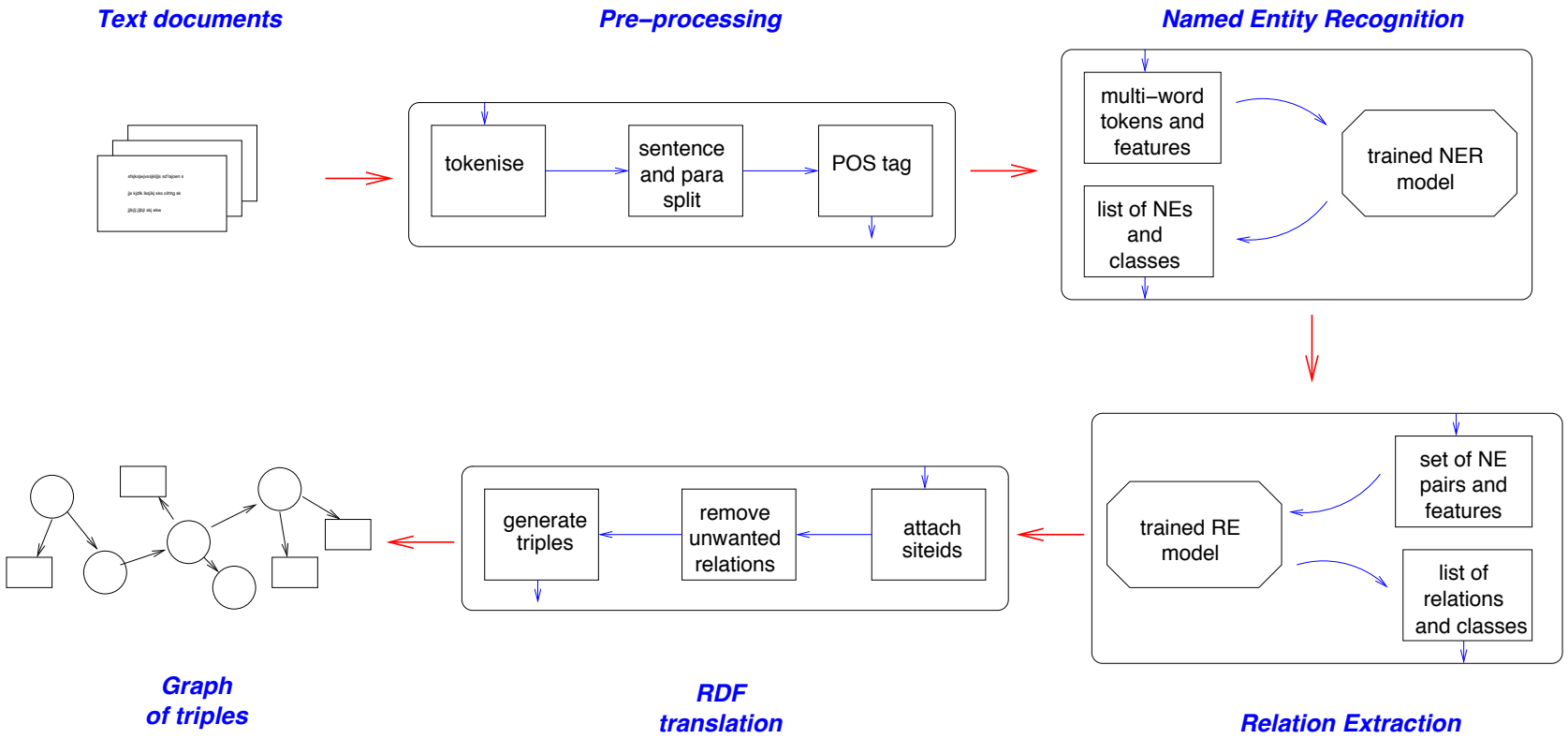
Relation	Avg Precision	Avg Recall	Avg F-score
eventAgent	97.46	82.18	88.72
eventAgentRole	0.00	0.00	0.00
eventDate	87.75	71.73	78.64
eventPatient	90.69	42.99	48.46
eventPlace	36.36	17.33	27.62
Overall Average	73.35	48.24	57.51

table source: authors presentation

Summary

- Event modeling is unorthodox in NER but results good
- Event relations are easier than others
- Extraction to RDF graph, as shown...
- Automatic extraction of events from text is feasible

txt2rdf



Img source: authors presentation

- Extra slides

- 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