Event Detection

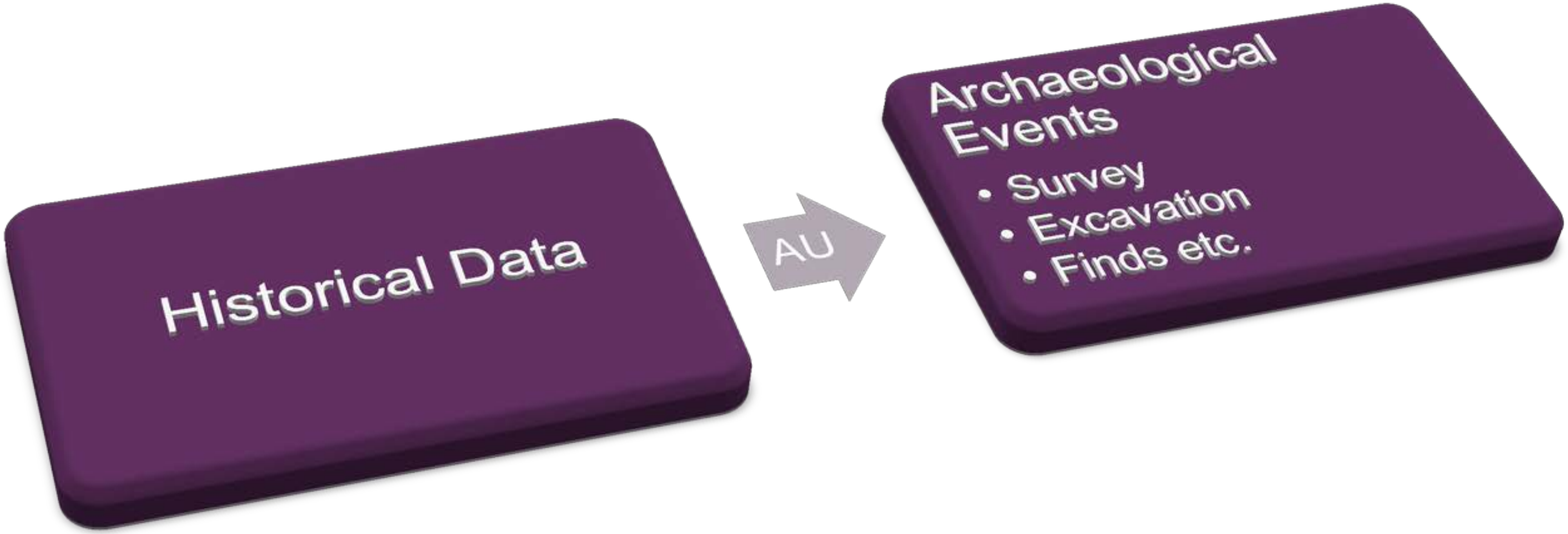
Automatic Extraction of Archaeological Events from Text

Wenbin Li
littletransformer@gmail.com
The Saarbrücken Graduate School of Computer Science

Outline

- Overview
- Background
  - Semantic web
  - Natural language processing
- Experiment
  - Settings (data)
  - Procedures
  - Results and evaluation (Remarks)
- Follow-ups

+ Overview *(what we do here)*



Historical Data

AU

Archaeological Events

- Survey
- Excavation
- Finds etc.

Background

- Semantic Web
- Natural Language Processing

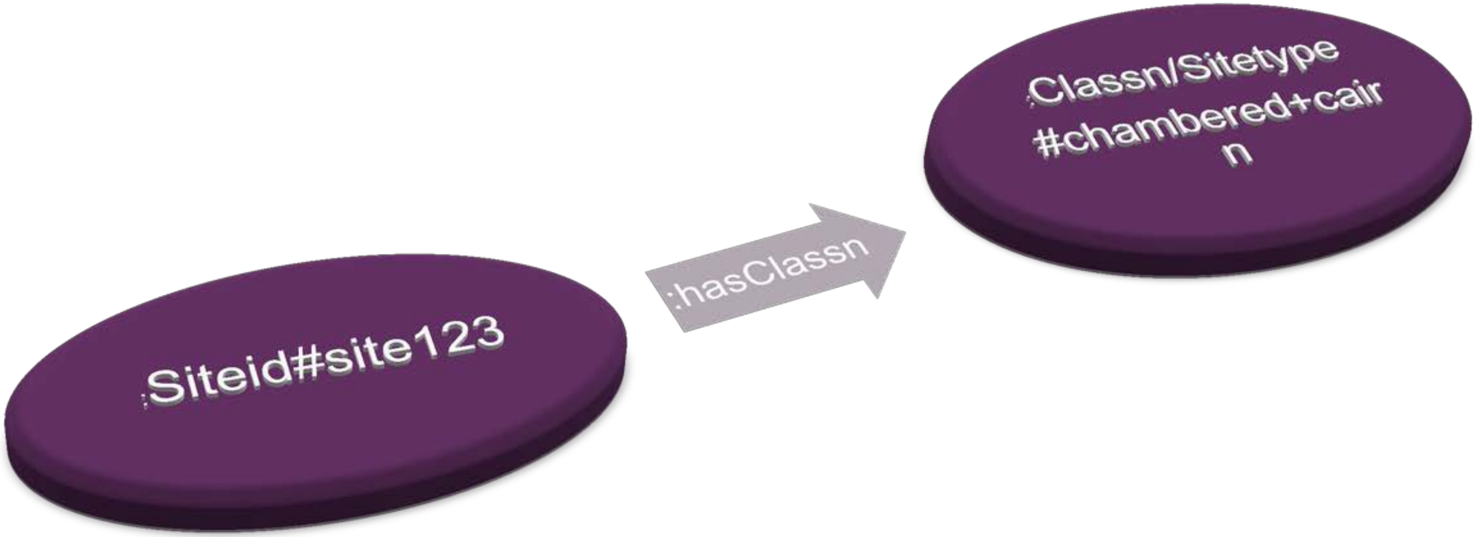
Semantic Web

- Reminder
  - A group of methods and technologies to allow machines to understand the meaning – or "semantics" – of information on the World Wide Web. (Wikipedia)

- RDF (Resource Description Framework)
  - A family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data(data about data) model. (Wikipedia)
  - RDF triple: subject-predicate-object
  - More information at http://www.w3.org/RDF/

Example of RDF triple

Statement: site123 is classified as a chambered cairn



Natural Language Processing

- Pre-processing
  - Tokenize
  - POS (Part-Of-Speech) tag
- NER (Name Entity Recognition)
  - Find and categorize the “entities” mentioned in a text
  - Typically include personal names, places, organization names and temporal expressions
- RE (Relationship Extraction)
  - Detect and classify semantic relationship from data

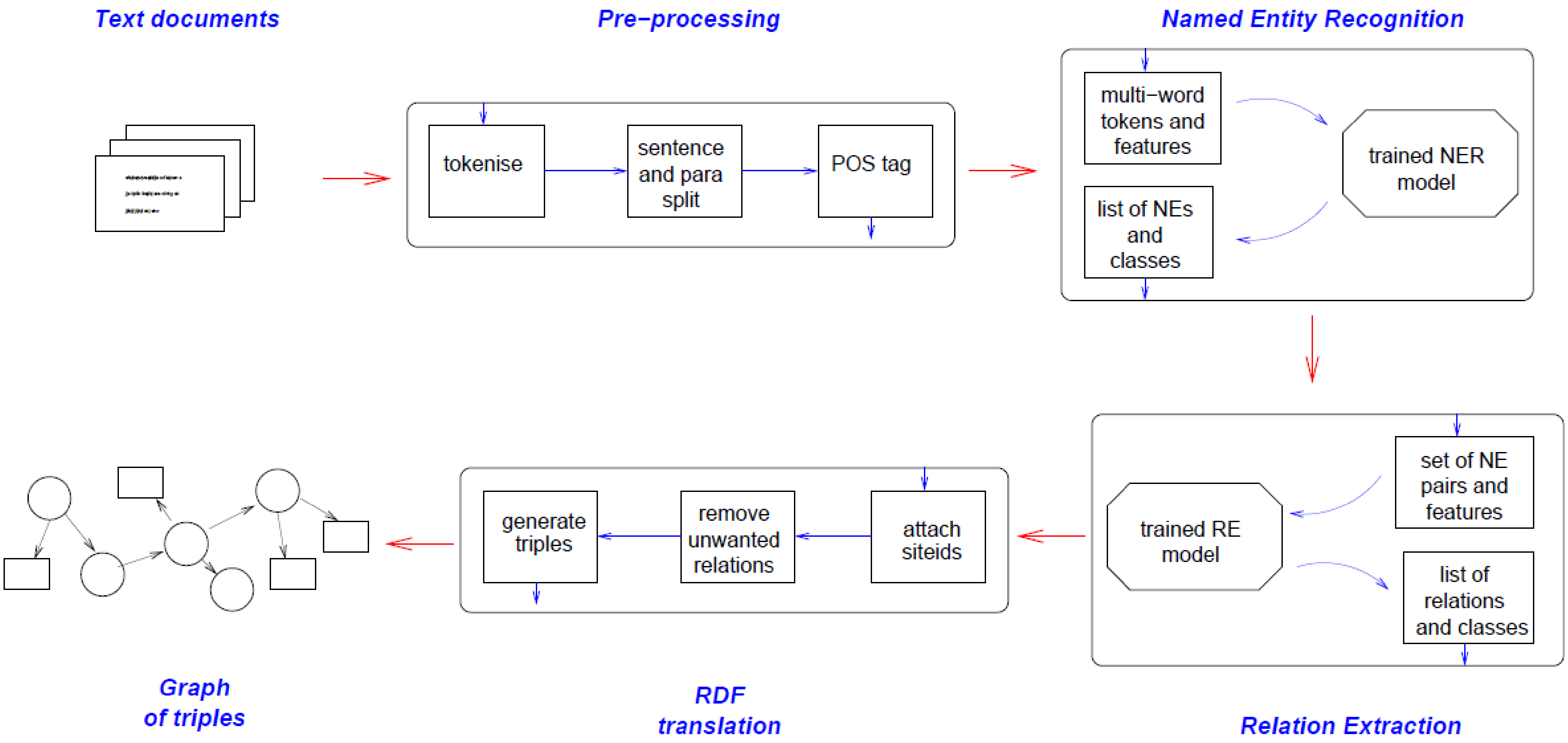
Experiment

- Data
- Procedures
- Evaluation

Data

- From RCAHMS (The Royal Commission on the Ancient and Historical Monuments of Scotland, http://www.rcahms.gov.hk)
- One of Scotland's 6 National Collection
- Recording Scotland's places, from the Neolithic to Now

Procedure


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
set of NE pairs and features
trained RE model
list of relations and classes
Relation Extraction
attach siteids
remove unwanted relations
generate triples
RDF translation
Graph of triples

+

Procedure--NER

- Supervised learning (training data ← hand-annotated documents)
- Domain specific classes

ORG	PERSNAME	ROLE	SITETYPE	ARTEFACT	PLACE
		√	√	√	
SITENAME	ADDRESS	PERIOD	DATE	EVENT	
√				√	

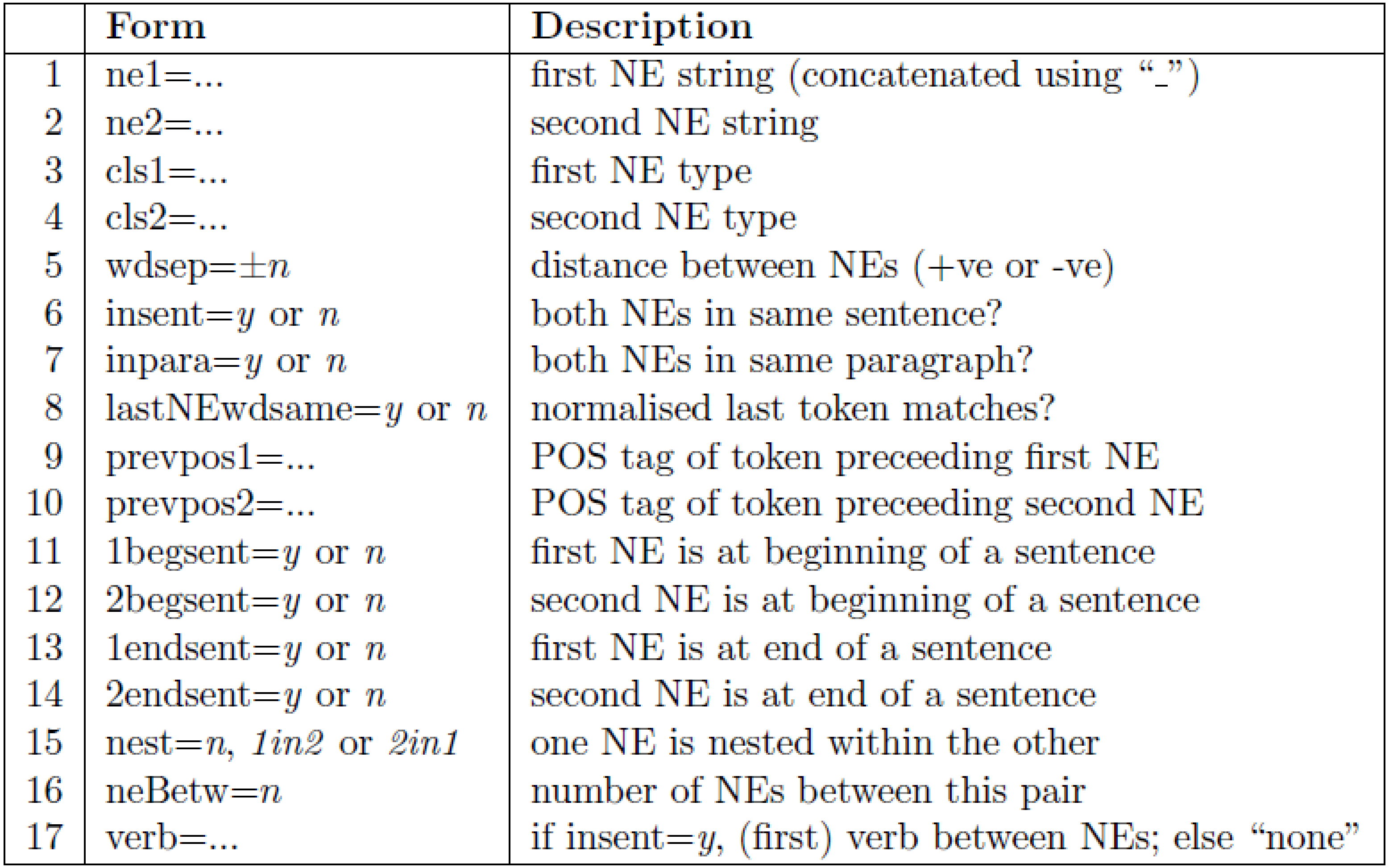
- NE nesting
  - [[[Edinburgh]^PLACE University]^ORG Library]^ORG

Procedure--RE

- Focus on event relationships
- Attributes of event
  - Agent
  - Role
  - Date
  - Patient
  - place
- Supervised learning (training data ← hand-annotated documents)

+

Learning process in NER & RE

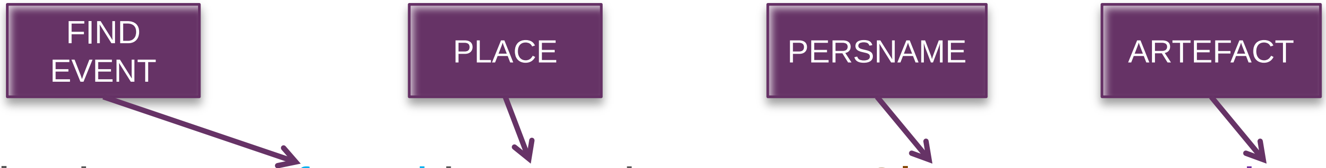


	Form	Description
1	ne1=...	first NE string (concatenated using "_")
2	ne2=...	second NE string
3	cls1=...	first NE type
4	cls2=...	second NE type
5	wdsep=$\pm n$	distance between NEs (+ve or -ve)
6	insent=y or n	both NEs in same sentence?
7	inpara=y or n	both NEs in same paragraph?
8	lastNEwdsame=y or n	normalised last token matches?
9	prevpos1=...	POS tag of token preceeding first NE
10	prevpos2=...	POS tag of token preceeding second NE
11	1begsent=y or n	first NE is at beginning of a sentence
12	2begsent=y or n	second NE is at beginning of a sentence
13	1endsent=y or n	first NE is at end of a sentence
14	2endsent=y or n	second NE is at end of a sentence
15	nest=n, *1in2* or *2in1*	one NE is nested within the other
16	neBetw=n	number of NEs between this pair
17	verb=...	if insent=y, (first) verb between NEs; else "none"

Procedure—Example

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

Procedure—Example(cont.)



FIND EVENT

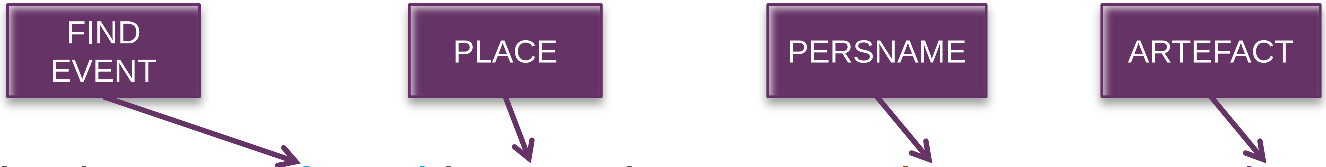
PLACE

PERSNAME

ARTEFACT

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

Procedure—Example(cont.)



FIND EVENT

PLACE

PERSNAME

ARTEFACT

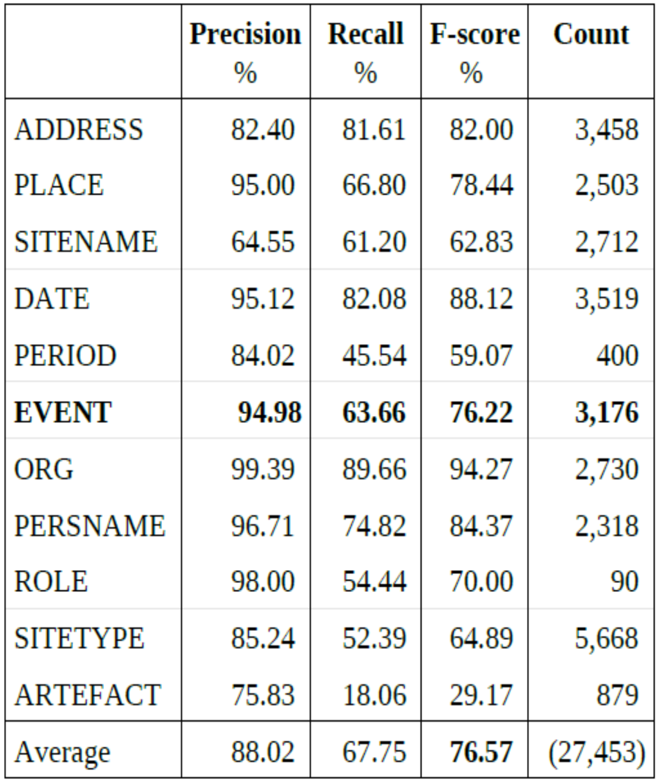
- The following were found in Unst by Mr A T Cluness: a steatite dish, …

Relationship	Entity1	Entity2
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

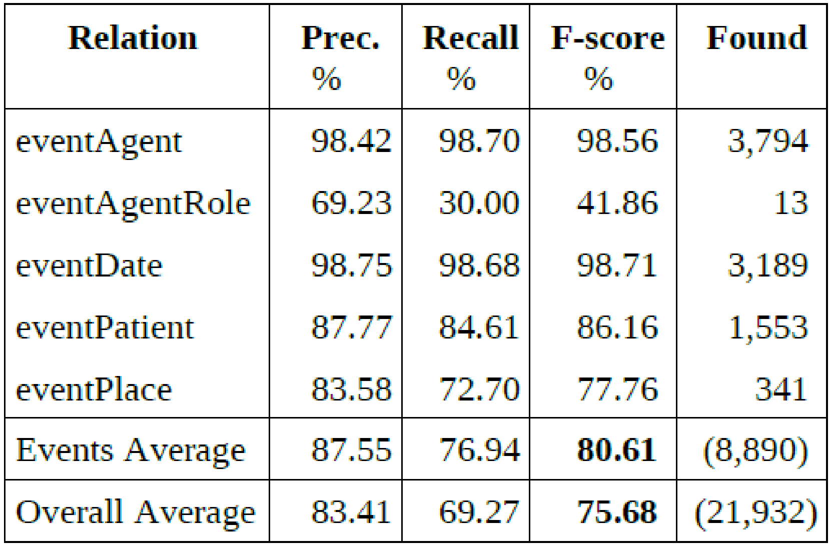
Evaluation

- NER evaluation
- RE evaluation
- NER and RE combination

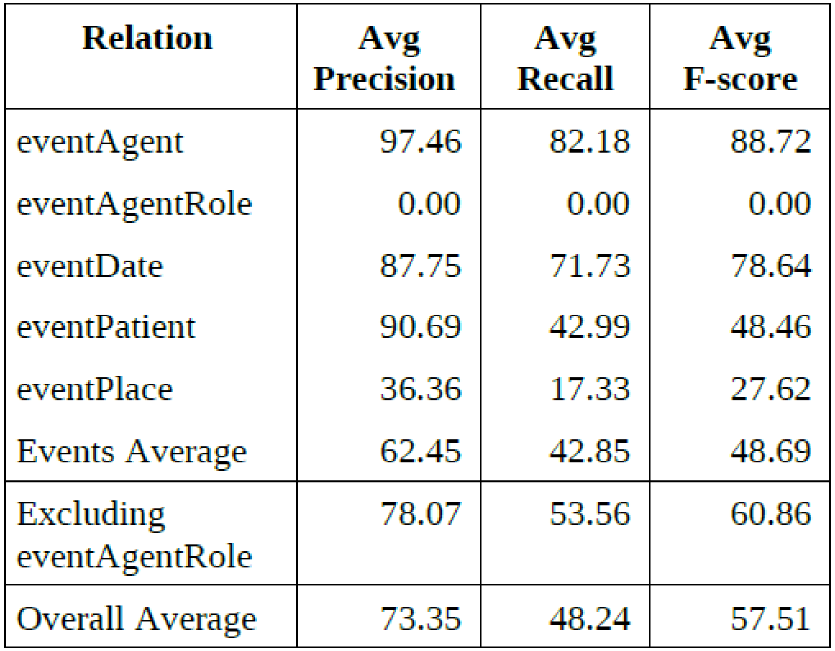
Some Results



	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)



Relation	Prec. %	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)



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
Events Average	62.45	42.85	48.69
Excluding eventAgentRole	78.07	53.56	60.86
Overall Average	73.35	48.24	57.51

Discussion of Results

- Weigh models towards preferring precision over recall
  - (?)when extracting facts from text, it more important to find correct statements than to find all that are available
- The author claims that the good results of eventAgent and eventDate in the pipeline suggests “with more data, the pipeline is capable of delivering very useful data structure without human labor”
  - (?)

Summary

- Practical application of NLP in event extraction in history domain

Extra: Follow-up Project

- Visualization

End

- Thank you!