Automatic Extraction of Archaeological Events from Text

Kate Byrne[1] and Ewan Klein[1]

[1]School of Informatics, University of Edinburgh. Scotland.

Abstract

This paper describes a series of experiments to automatically detect and categorize archaeological events—such as survey, excavation, finds and so forth—that are described in natural language text documents. Complex event structures with attributes including date, agent, and location are extracted and converted into families of binary relations. These in turn can be mapped to RDF triples for publication as Semantic Web graphs, with the potential of making it dramatically easier to interconnect separate data silos. We present results indicating that although events do not conform to the standard definitions of "entities", they can be detected with high precision, making large-scale processing of text documents a practical possibility.

Keywords: *Semantic Web, RDF, relation extraction, cultural heritage, event-based recording*

1   INTRODUCTION

Archaeological data—indeed, cultural heritage information in general—is typically managed in hybrid structures combining structured database fields, free text documents, and masses of supporting archives in various formats: maps, photographs, measured survey plans, and so forth. There have been rapid advances in information management in the last decade or two, allowing these multimedia resources to be exploited in exciting ways, but the greatest challenge still remains: to deal really effectively with the natural language text. We still lack robust mechanisms for working out what a piece of text is about, what facts it expresses and what happenings it relates. Yet natural language is, of course, humankind's favoured way of conveying information.

This paper deals with a specific example of the problem as just posed. The data used in this work comes from the National Monument Record of Scotland (NMRS) maintained by RCAHMS,[1] and consists of a relational database based around site records, where each site has a set of text notes associated with it. The text documents generally describe a particular site in terms of professional visits made to it, surveys, excavations, and so on. For example, the following text excerpt from a site record in the NMRS database (site HP60NW 3) concerns a number of excavation finds made at the site and now located in the Shetland Museum:

> *The following were found in Unst by Mr A T Cluness, Ross-park, Uyeasound. Fragments of four steatite vessels (ARC 65516), fragment of steatite dish (ARC 65515), handled club (ARC 65514), hammerstone or*
>
> *pounder (ARC 65512). Shetland Museum accessions register.*

This text tells us that four artifacts or sets of artifacts were found, at Unst, by Mr. A. T. Cluness. As will be explained, we treat this as four "find events", each related to the same site and having the same location and agent but with a different "patient" (the thing found) in each case. There are other facts for our system to extract, such as the relationship between Mr. Cluness and Ross-park (his residence) and between Ross-park and Uyeasound (a locational relationship), but in this paper we concentrate on event relations. These exemplify our methods and are of particular interest because event-based rather than purely site-based recording is becoming widespread in historic monument data management. We return to this example in Section 5, where we discuss the details of how event statements are detected. The statements are extracted as sets of related subject-predicate-object triples, such as:

```
find1-hasLocation-Unst
find1-hasAgent-A_T_Cluness
```

These can then be turned into RDF or other formats as desired. These two simple statements are represented graphically in figure 1.
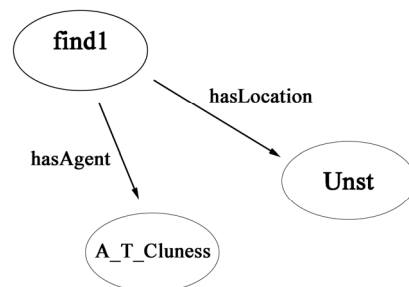

Figure 1. *Graphical representation of triple statements.*

We aim to extract the "events" using Natural Language Processing (NLP) tools that can operate over very large batches of text. Each event becomes a data structure with attributes such as date, agent, and location, and

---

these slots will be automatically given values if the information is present in the text. We use the RCAHMS data for practical experimentation, but the methods are designed to be generic for the cultural heritage domain.

The potential for exploiting the resulting event structures is considerable. They could, for example, be used to populate relational database (RDB) tables for subsequent SQL querying. See Sporleder et al.[1] for a study of RDB population in a natural history domain, based on Information Extraction techniques that have some similarities to our methods.

The focus in our work is on transforming event structures into RDF graphs that can be integrated—along with RDB data—into the Semantic Web. The opportunities this provides are described in the following section. In Section 3 we give an overview of the NLP "pipeline" that takes in plain text at one end and produces an RDF graph at the other. The details of the key NLP components, Named Entity Recognition (NER) and Relation Extraction (RE) are explained in Sections 4 and 5, and the transformation of text relations into RDF is covered in Section 6. The last two sections show formal evaluation results using the standard metrics of the field and discuss our overall conclusions.

2 THE SEMANTIC WEB

The Semantic Web is steadily gathering momentum and, after a slow start lasting almost a decade, we may now be on the brink of explosive growth similar to that of the original Web. With Google recently announcing support for embedded RDFa in web pages,[2] the use of RDF is moving out of the academic ivory towers and into the mainstream. If the new "Data Web" supersedes the existing "Document Web" it is vital that the cultural heritage material curated in archives around the world becomes part of it. See Schreiber et al.[3] and Hyvönen et al.[4] for example initiatives.

One of the advantages of the Semantic Web is the capacity for effortless interconnection with related information that is already in RDF format. If the new graph one creates has any nodes in common with the wider graph, they are automatically connected. For example, we translated the thesauri used by RCAHMS—which are based on the MIDAS Heritage standards[5]—into RDF using a normalized URI format, so that technical terms mentioned in the text are automatically "grounded" against the relevant thesaurus, with their broader, narrower and related terms available and an explanatory scope note. Figure 2 shows a specific example. A simple RDF triple expressing the statement "site123 is classified as a chambered cairn" can be represented, in RDF subject-predicate-object format, as:

If this fact were extracted in isolation from text, it would inherit all of the grounding structure shown in figure 2. No extra connections need be made, and the whole panoply of thesaurus information is instantly at the disposal of subsequent SPARQL queries. As the figure indicates, the SKOS framework[6] was used to encode the thesaurus structures.

It has been shown[7] that the MIDAS Heritage thesauri can in turn be integrated with the CIDOC-CRM,[8] allowing yet wider conceptual grounding. Similar connections can be made to any related SemanticWeb data, perhaps from related cultural archives. For example, a common graph representation would allow site records to be integrated with museum finds, on shared nodes (such as find location).

3 OVERVIEW OF THE *txt2rdf* PIPELINE

The experiments described in this paper were part of a larger project on populating the Semantic Web by combining RDB data with structured relations extracted from text using NLP methods. Figure 3 shows the overall layout of the system, which was named *Tether*.

[1]Caroline Sporleder et al., "Cleaning and Enriching Research Data on Reptiles and Amphibians. The MITCH Pilot Project and 'Nulmeting'," *Technical Report ILK 06-01* (Tilburg University, February 2006).

[2]Posting on the Google Webmaster Central blog, 12 May 2009, http://googlewebmastercentral.blogspot.com/2009/ 05/introducing-rich-snippets.html.

[3]Guus Schreiber et al., "MultimediaN E-Culture Demonstrator," in *Proceedings of the International Semantic Web Conference* (*ISWC 2006*) vol. 4273 (Athens, GA: LNCS, 2006) 951–958.

[4]Eero Hyvönen et al., "CultureSampo—Finnish Culture on the Semantic Web: The Vision and First Results," in *Information Technology for the Virtual Museum*, ed. K. Robering (Berlin: LIT Verlag, 2007) 25–26.

[5]Edmund Lee, ed., *MIDAS HeritageA Data Standard for the Historic Environment*, Forum for Information Standards in Heritage (FISH) 2007.

[6]www.w3.org/2004/02/skos/.

[7]Ceri Binding, Keith May and Douglas Tudhope, "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM," in *Proceedings of European Conference on Digital Libraries (ECDL08)* (Aarhus, Denmark: Springer-LNCS, 2008) 280–290.

[8]Nick Crofts et al., eds. *Definition of the CIDOC Conceptual Reference Model* (CIDOC, March 2008, ed. 4.2.4). ISO 21127: 2006.
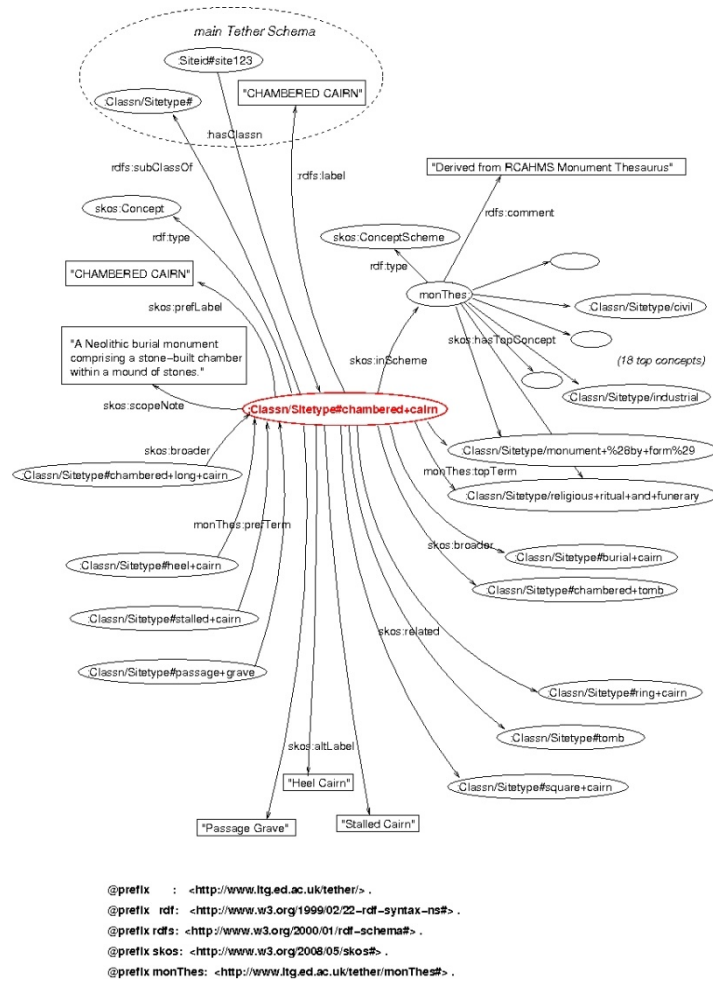
Figure 2. *Example illustrating the grounding of terminology against a pre-existing graph. (Compact URIs are used for brevity, with prefixes are listed at the bottom of the figure).*

The event extraction component is part of the *txt2rdf* pipeline—a sequence of NLP procedures in which the results of each step are passed into the next step. This pipeline starts with plain text documents as input and the final output is a graph of RDF triples. As illustrated in figure 4, the pipeline starts with some standard pre-processing steps: splitting the text into "tokens" (which correspond approximately with words), finding the sentence and paragraph breaks, and then doing shallow parsing to annotate each token with a tag indicating its part of speech (POS). Once these basic preparatory steps have been done the key procedures can be carried out: Named Entity Recognition (NER) followed by Relation Extraction (RE). These are described in the next two sections. The final operation is to transform the relations into RDF triples and anchor them to individual sites from the RDB data. This involves generating suitable URIs for all the graph nodes and edges, because being accessible to HTTP is fundamental to Semantic Web data structures.

The pipeline is designed to identify a number of binary relations between pairs of named entities (see Section

4) including *partOf*, *sameAs*, *seeAlso*, *instanceOf*. Here we focus particularly on the event relations: *eventAgent*, *eventAgentRole*, *eventPatient*, *eventDate* and *eventPlace*, which are described further in Section 5 below.
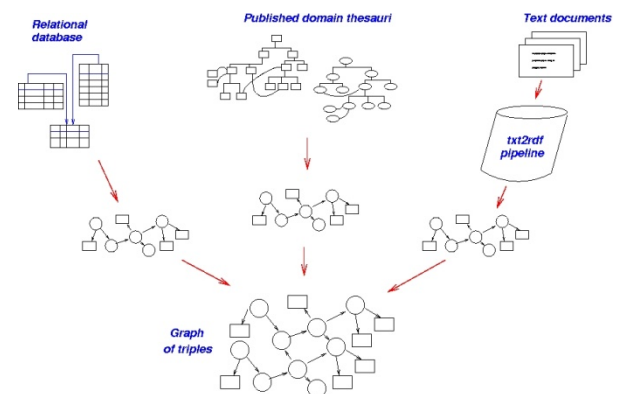

Figure 3. *An overview of the Tether system. See fig. 4 for the txt2rdf pipeline.*
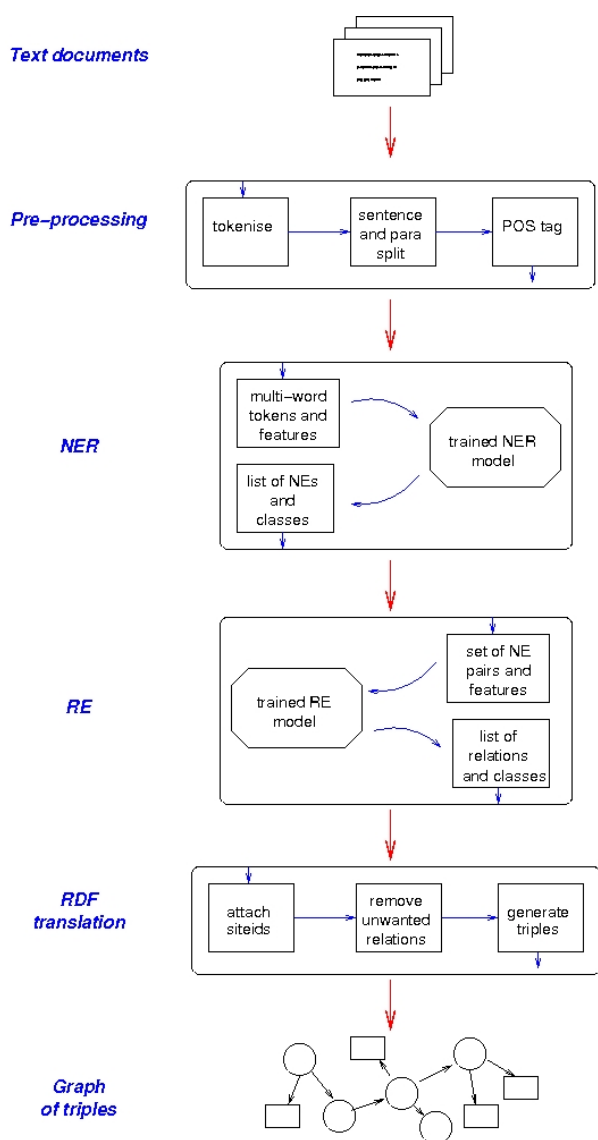
Text documents

Pre-processing

NER

RE

RDF translation

Graph of triples

Figure 4. *The text to RDF pipeline* (txt2rdf).

4 NAMED ENTITY RECOGNITION

Traditional NER involves finding and categorizing the "entities" mentioned in a text. These are noun phrases that can be loosely characterized as "content carrying terms"; they typically include personal names, places, organization names and temporal expressions. For this domain we include new classes: site name, site type, and object type. We also require a degree of granularity in the spatial designations, as locational information is of particular importance in site-based recording. The 11 separate NE classes are listed in the NER results table in Section 7. The calculation of scores follows normal practice for the NER task, as explained in Section 7. Various conferences have evaluated NER systems in shared task competitions, in particular MUC[1] and CoNLL.[2] The CoNLL 2002 and 2003

competitions are particularly good sources of information (see, for example, Malouf,[3] and Curran and Clark[4]). The "CandC" system[5] used in our experiments was developed for CoNLL-2003.

As the first step towards extracting event structures, we designate an additional NE class for EVENT, which includes site visits, excavations, surveys and so forth. (In all, the subclasses of EVENT are: SURVEY, EXCAVATION, FIND, VISIT, DESCRIPTION, CREATION, and ALTERATION. The first five are the ones most frequently encountered in the RCAHMS data.) Mentions of events within the text are very often through verb phrases: "site X *was visited* on [a date]...", "site Y *has been recorded* by [an agent]". Our results show that treating events as reified entities leads to high performing extraction, even though they are not Named Entities as usually construed.

The attributes of an event (where and when it took place and who was involved) take their values from other NE classes (such as PLACE, DATE, PERSNAME). The event graph structure will emerge as a collection of binary relations between a specific event and its attribute values.

The techique used to find entity mentions in the text documents is supervised machine learning. This involves building a mathematical model of what NEs from each category are like, based on a set of training examples that have to be prepared in advance. Once the model has been trained it can be used to classify fresh texts from the domain.

Entity mentions can be nested within each other. In the RCAHMS corpus, up to three levels of nesting can occur. For example, in the string

[[[Edinburgh]PLACE University]ORG Library]ORG

the token "Edinburgh" is a PLACE entity mention on its own, and part of two distinct organization (ORG)

[1]Message Understanding Conference, www-nlpir.nist.gov /related_projects/muc/.

[2] Conference on Natural Language Learning, www.ifarm.nl/signll/conll/.

[3]Robert Malouf, "Markov Models for language-independent named entity recognition," in *Proceedings of CoNLL-2002*, ed. Dan Roth and Antal van den Bosch (Taipei, Taiwan, 2002) 187–190.

[4]James R. Curran and Stephen Clark, "Language Independent NER using a Maximum Entropy Tagger," in *Proceedings of CoNLL-2003*, ed. Walter Daelemans and Miles Osborne (Edmonton, Canada, 2003) 164–167.

[5]James Curran and Stephen Clark, "Maximum Entropy Tagging for Named Entity Recognition," *ICCS: Informatics research report* (University of Edinburgh, School of Informatics, 2003).

entity mentions. In contrast to much of the work in the NER field, we pay particular attention to nesting because there is so often a relationship between inner and outer entity mentions. In the example just given, "Edinburgh" is the location of the two organisation entities, and the library is part of the university. The methodology is described in Byrne.[1]

5 EXTRACTION OF EVENT RELATIONS

Interest in the textual relation extraction problem has grown considerably over the last few years. The NIST-sponsored ACE (Automatic Content Extraction) program[2] has been running since 2000, with research goals of detecting and characterizing entities, relations, and events. The ACE tasks are complex and include many more characteristics of text relations than are needed for the purposes of *Tether*, where the aim is to find pairs of related entities and label the relationship between them. Here we focus particularly on event relationships.

The attributes of an event are defined as being the *agent* who was responsible for the event taking place, the *role* of that agent (surveyor, sponsor, or whatever), the *date* on which it occurred, the *patient* (the thing that experienced the event, such as an artifact being the patient of a finding event), and the *place* where the event happened. The graph structure produced by the *txt2rdf* pipeline for each event takes the form of a set of binary relations between pairs of NEs, where the subject NE in each case is an instance of the EVENT class and the object is an instance of one of the relevant classes: ORG or PERSNAME for the agent, ROLE for the agent's role, and so on. These binary relations are derived from a complex event relation with parameters *eventAgent*, *eventAgentRole*, *eventPatient*, *eventDate,* and *eventPlace*. In any given text, there may only be values for a subset of these slots—for instance, the date of an event may be mentioned but not the agent.

The Relation Extraction (RE) task is, like the NER step, treated as a supervised classification exercise using hand-annotated training documents in which textual relations have been marked. The classifier is presented with a list of all possible pairings of NEs from the text, along with a set of "features" that give clues to whether the pair are likely to be related or not. The training run involves converting these clues into a probabilistic model of the characteristics of a related pairing. Once trained, the classifier can label new NE pairings, by comparing their features with its model.

For each fresh NE pairing presented, the classifier will determine whether the two NEs are related and, if they are, what the relationship type is. Figure 5 (p. 54) illustrates how the event detection process works. The text fragment contains four entity mentions, identified and categorized by the NER step of the pipeline.

A list of the six possible pairings is generated, and against each pairing the first few features are shown on the right. The left-most label is the relation classification. An "O" is the conventional way of indicating that there is no relationship. Where there is a relationship between the pair, its category label is given. Thus the first line shows that there is an *eventLocation* relation between the FIND EVENT (whose mentioning string is "were found") and the PLACE entity "Unst". A set of 17 features was used, including the category of each NE in the pairing, their distance (in words) from each other in the text, their POS tags, whether one is nested within the other, and so on. The full list is given in table 1.[3]

	Form	Description
1	ne1=...	first NE string (concatenated with "_")
2	ne2=...	second NE string
3	cls1=...	first NE type
4	cls2=...	second NE type
5	wdsep=±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	lastNEwordsame=y or n	normalized last token matches?
9	prevpos1=...	POS tag of token preceding first NE
10	prevpos2=...	POS tag of token preceding second NE
11	1begsent=y or n	first NE is at beginning of a sentence
12	2begsent=y or n	second NE is a 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*

Table 1. *Textual features used for building RE model.*

6 MAPPING RELATIONS TO RDF

As explained, finding textual relations is essentially a two-phase procedure: the NE mentions are detected

---

[1]Kate Byrne, "Nested Named Entity Recognition in Historical Archive Text," in *Proceedings of ICSC2007, IEEE International Conference on Semantic Computing* (Irvine, California, Sept., 2007).

[2]www.nist.gov/speech/tests/ace/, www.ldc.upenn. edu/Projects/ ACE/.

[3]Some extra features based on subclass labels were also used but they are of secondary importance. See Kate Byrne, *Populating the Semantic Web—Combining Text and Relational Databases as RDF Graphs* (Ph.D. diss., University of Edinburgh, 2009) chapter 8.

first and then relations are looked for between pairs of them.

The upper part of figure 6 shows a sample text (for site456) in which the NE strings are highlighted and enclosed in square brackets. (The display is from the MMAX2[1] annotation tool.) In this example text, the word "recorded" is taken as the EVENT mention for six separate surveying events (which is why it is surrounded by six sets of brackets) that are listed in PLACE- SITETYPE pairs later in the text. One of the six is highlighted, with an *eventPatient* of "Sub-rectangular cairn" (the thing that was recorded) and an *eventPlace* of "ND 3342 8884" (the Ordnance Survey grid reference for the location of the cairn). The lower part of figure 6 indicates how this complex event relation is split into binary relations, expressed as `subject-predicate-object` triples, ready for the next conversion step to RD. The event has first to be grounded within the wider graph by linking it to its parent site with a *hasEvent* predicate. In this illustration, the survey event is labeled "recording" to identify it uniquely;[2] it must, for example, be distinguished from the other five recording events that have different location and patient properties. Once this unique identifier is established, it is straightforward to add the necessary binary relations—in this case *hasLocation* and *hasPatient*—for the remaining event properties.

The final step in mapping to RDF is to convert these basic triples into valid RDF, with suitable URIs to identify each property and object "resource" (to use RDF terminology).

At this point a design decision is required about whether the object of the triple should be a resource with a full URI, or a typed literal (such as a quoted string). In the *Tether* design all the values illustrated in figure 6 become resources with their own URIs, and are therefore able to act as the subject of other triples that may subsequently be added to the graph. (In RDF the subject of a triple must be a resource, not a literal.) The form of the URI is to some extent arbitrary—as long as it uniquely identifies the correct resource it can be anything the designer chooses—but in Tether the generated URIs are based on a normalized version of the original text string, partly to aid human readers.

In addition to the three "statements" extracted from the text in this example, further RDF triples are needed to locate the event correctly in the wider RDF schema. Each resource node is typed (using an *rdf:type* predicate) to show which class of objects it belongs to. The node derived from "Sub-rectangular cairn" will be

a member of the "sitetype" class, for instance, which in turn is a subclass (*rdfs:subClassOf*) of the "classification" class. This enables later graph queries to, say, find all surveys of a particular kind of site.

7   RESULTS AND EVALUATION

As is usual in NLP classification experiments, the NER and RE steps were evaluated against the hand-annotated gold standard, measuring precision and recall and calculating an F-score from them.[3] Table 2 (p. 55) summarizes the results for the NER step, highlighting the EVENT class in particular. Similar kinds of NE classes are grouped together—for example, ADDRESS, PLACE, and SITENAME all contain locational information. A less granular system (for example, lumping these three together) would be likely to achieve higher scores. Some of the classes (such as ROLE and PERIOD) are very sparsely populated and their results should be treated with caution, as it is impossible to be sure that the RCAHMS data can provide a representative model of these entity types. The overall F-score across all NE classes was 76.57% and the EVENT class score is very close to this (76.22%), indicating that EVENTS are no more difficult than the average to detect despite the fact that, as explained in Section 4 above, the way we model them is anomalous in standard NER terms.

Contrary to standard practice, we deliberately weighted our models towards preferring precision over recall, for both the NER and RE steps. Our argument is that, when extracting "facts" from plain text, it is far more important to find correct statements than to find all that are available. The natural redundancy of normal language will tend to ensure that really important facts are not missed by this tactic. (As the Bellman said, "What I tell you three times is true."[4]) Conversely, a system that is known to produce a high proportion of incorrect facts will be difficult to trust. Particularly if the data is used to populate Semantic Web graphs that will be available to a very wide and non-expert user base (as we hope), it seems very important not to poison the well with false information. It should be noted that the precision for EVENT detection is very good (94.98%).

The RE results that pertain to event relations are given in table 3. These figures are for the RE step in isolation, when relations are extracted over gold standard NEs. As mentioned in Section 3, the *txt2rdf* pipeline extracts a number of other relations as well as

---

[1]www.eml-research.de/english/research/nlp/download/ mmax. php.

[2] In the actual *Tether* implementation, unique URIs were generated using document and words IDs.

[3]Precision is the fraction of all output results that were correct, and recall is the proportion of the entire correct population that was found. The F-score (sometimes "F1 score" to distinguish it from variants) is the harmonic mean of precision (*P*) and recall (*R*), *viz. 2PR/(P+R)*.

[4]From *The Hunting of the Snark* by Lewis Carroll.

the event ones that concern us here. The average score over all relations (75.68% F-score) is shown for comparison with the events average (80.61%)—it is encouraging to note that these event relations, which we consider to be especially important for retrieval applications, are actually easier than others to find. An indication of the size of each relation set is given in the table and, as with NE classes, the scores for the smaller sets will be unreliable. The *eventAgentRole* relation set is vanishingly small and should probably be discounted altogether. (The average score would of course be higher if it were discounted.)

Finally, table 4 lists the results for the NER and RE steps in combination, i.e. relation detection over automatically extracted NEs, which are not all correct. The comparison is with the gold standard and the datasets used for training and testing are necessarily smaller. The test set that the scores were calculated over is only 10% of those used for the separate NER and RE evaluations. (In those evaluation steps, 10-fold cross validation was used, which enables the entire annotated corpus to be used for both training and testing.) One result of this is that the smaller categories, like *eventAgentRole,* effectively disappear altogether. Two averages are therefore shown for events, the second excluding *event AgentRole*. This result (78.07% precision and 60.86% F-score) compares with the overall average for all relation types (73.35% precision and 57.51% F-score) in much the way one would expect from the earlier results.

For practical purposes, an important finding from our experiments is that the two largest categories (*eventAgent* and *eventDate—*see table 3) both show good or (in the case of *eventAgent*) very good results, not just for precision, which we are weighting the model towards, but also for recall. (Clearly if we can get good recall without losing precision this is desirable.) This suggests that, with more training data or models otherwise made more accurate, the *txt2rdf* pipeline is capable of delivering very useful data structures without human labor. Once the initial models have been created, the overhead of processing almost any volume of data is negligible.
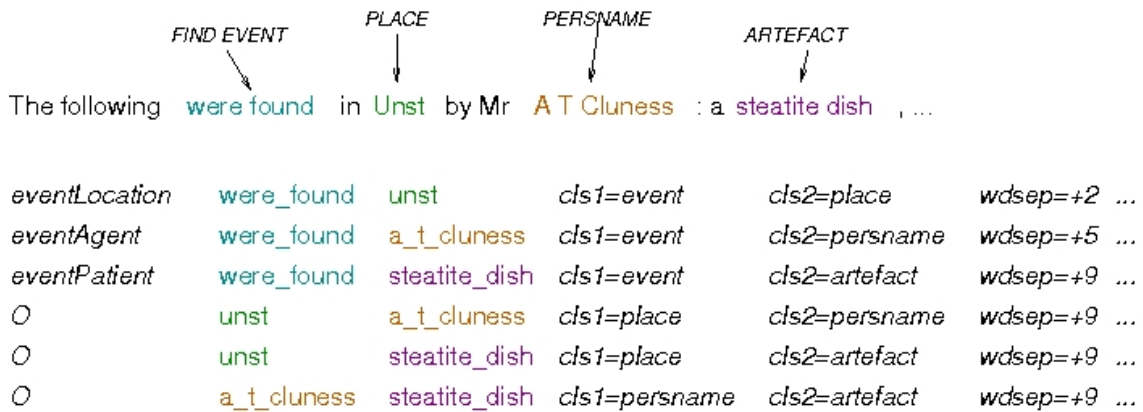

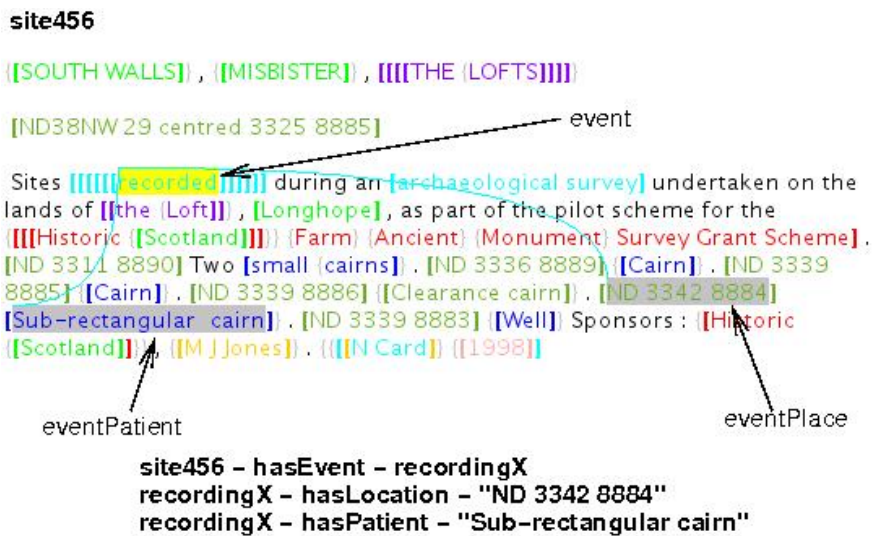**Figure 5.** *Supervised learning for event extraction.*

Figure 6. *Mapping relations to RDF.*

	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
ARTIFACT	75.83	18.06	29.17	879
Average	88.02	67.75	**76.57**	(27,453)

Table 2. *Summary of NER results.*

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)

Table 3. *Summary of RE results for 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
Events Average	62.45	42.85	48.69
Excluding eventAgentRole	78.07	53.56	60.86
Overall Average	73.35	48.24	57.51

Table 4. *Results for whole pipeline, NER and RE combined.*

8   CONCLUSIONS

We have shown that, although our method of modeling events as reified entities may be unorthodox in NER terms, the results justify it, as recognition scores for instances of the EVENT class (precision 94.98% and F-score 76.22%) are close to or better than the average we achieved for the domain (precision 88.02%, F-score 76.57%). Moving on to detecting relations between events and hence extracting a structured subgraph of triples, the results for event relations (precision 87.55% and F-score 80.61%) are several percentage points better than the average across all text relations attempted (precision 83.41%, F-score 75.68%). We argue that precision should be preferred to recall for this particular information extraction task, where accuracy is more important than complete coverage, and our models are weighted accordingly. For the combined pipeline of tasks our measured precision for event extraction is 78.07% (though at the expense of a low recall score).

The results are encouraging, as event extraction is an important task for cultural data management and appears to be easier than finding other textual relations (such as "part of", "instance of" and so on). Event based recording is widespread in archaeological and historical site management, but is often locked into free text notes. We show that automatic discovery of structured events is a practical possibility. With precision around the 80% mark, there is still a need for human intervention, but the "easier" cases can be determined by automated NLP techniques, potentially saving enormous amounts of skilled investigator time.

Once extracted, the event structures—with properties such as agent, date, location, and so on—can be used either to populate traditional database tables or, as suggested here, to build Semantic Web graphs whose intrinsic purpose is to facilitate the linking of related but separate datasets. Fully searchable integrated cultural archives, for instance linking museum finds to archaeological sites, become a feasible proposition using Semantic Web tools. While previous work in this area has generally used relatively small pilot data collections, our results suggest that, through the use of domain-trained models, there is no reason why full-scale conversion of complete archives should not be attempted.

BIBLIOGRAPHY

Binding, C., K. May, and D. Tudhope. "Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM," in *Proceedings of European Conference on Digital Libraries* (ECDL08), no. 5173 (Aarhus, Denmark: Springer–LNCS, 2008), 280–290.

Byrne, K. "Nested Named Entity Recognition in Historical Archive Text," in *Proceedings of the IEEE International Conference on Semantic Computing* (ICSC2007), Irvine, CA, 2007.

Byrne, K. *Populating the Semantic Web—Combining Text and Relational Databases as RDF Graphs*. Ph.D. diss., University of Edinburgh, 2009.

Crofts, N., M. Doerr, T. Gill, S. Stead, and M. Stiff, eds. *Definition of the CIDOC Conceptual Reference Model*. CIDOC, 4.2.4 edition, 2008. ISO 21127: 2006.

Curran, J. R., and S. Clark. "Language Independent NER using a Maximum Entropy Tagger," in *Proceedings of CoNLL-2003*, edited by W. Daelemans and M. Osborne, 164–167. Edmonton, Canada, 2003.

Curran, J., and S. Clark. "Maximum Entropy Tagging for Named Entity Recognition," *Informatics Research Report*, University of Edinburgh, School of Informatics, ICCS.

Hyvönen, E., T. Ruotsalo, T. Häggström, M. Salminen, M. Junnila, M. Virkkilä, M. Haaramo, E. Mäkelä, T. Kauppinen, and K. Viljanen. "CultureSampo–Finnish Culture on the Semantic Web: The Vision and First Results," in *Information Technology for the Virtual Museum*, edited by K. Robering, 25–36. Berlin, LIT Verlag, 2007.

Lee, E., ed. *MIDAS Heritage—A Data Standard for the Historic Environment*. Forum for Information Standards in Heritage (FISH), 2007.

Malouf, R. "Markov Models for Language-independent Named Entity Recognition," in *Proceedings of CoNLL-2002*, edited by D. Roth and A. van den Bosch, 187–190. Taipei, Taiwan, 2002.

Schreiber, G., A. Amin, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, L. Hollink, Z. Huang, J. van Kersen, M. de Niet, B. Omelayenko, J. van Ossenbruggen, R. Siebes, J. Taekema, J. Wielemaker, and B. Wielinga, "MultimediaN E-Culture Demonstrator," in *Proceedings of the International Semantic Web Conference* (ISWC 2006), vol. 4273. Athens, Georgia: LNCS, 2006, 951–958.

Sporleder, C., M. van Erp, T. Porcelijn, A. van den Bosch, P. Arntzen, and E. van Nieukerken. "Cleaning and Enriching Research Data on Reptiles and Amphibians. The MITCH Pilot Project and 'Nulmeting'," *Technical Report ILK 06-01*, Tilburg University, 2006.