MASWS
Natural Language and the Semantic Web

Kate Byrne

School of Informatics

14 February 2011
Populating the Semantic Web
why and how?
text to RDF

Organising the Triples
schema design
open issues and problems

References
The Semantic Web needs data

- Semantic Web first proposed in 1999
- Why hasn’t it taken off the way WWW did?

not enough data

queries aren’t useful

I’ll delay adding my data
How to get data

How do we populate the Semantic Web?

create
structured data

convert
unstructured and semi-structured
Semi-structured content is everywhere
Why convert unstructured text?

- Extracting RDF triples from free text is not easy
- So why is it worth doing?
  - there was knowledge before WWW (!)
  - high quality material, often historical
  - big digitisation push in 1990s (and again in 2011?) –
    - – but still challenging to search text effectively
  - natural language is our preferred way of communicating
- Can the web learn to talk in natural language...?
“We’ll start out by speaking in simple declarative sentences.”
Natural Language Processing

- Extraction of semantic content from natural language text
- Systems now available – Open Calais, Powerset, AlchemyAPI
- To explain the process: part of my PhD work

"Kate likes chocolate."

Think of RDF triples as declarative sentences:
RDF nodes = nouns; RDF arcs = verbs

prefix : <http://www.ltg.ed.ac.uk/>
Converting text to RDF

fancy NLP processing and RDFisation
Evidence of a quartz knapping site was found within the confines of the [stone circle][SITETYPE], and in conjunction with several structures within the inner ring, strongly suggests a domestic site.

Besides the quartz implements and corresponding waste, several other artifacts of local origin occurred including a split pebble axe of greenstone with [Shetland][PLACE] [Early Bronze Age][PERIOD] affinities. [B Beveridge][PERSNAME], [1972][DATE]. Field survey and [excavation][EVENT], as a response to continual wind and marine erosion, was carried out at the [Sands of [Breckon][ADDRESS]][SITENAME] between [1982][DATE] and 1982 [HP50NW 11.00].

Occupational debris from the early [Iron Age][PERIOD] was discovered in the surrounding deposits of an undated structure. The structure was surrounded by debris of an undated structure. The structure was surrounded by debris of an undated structure. The structure was surrounded by debris of an undated structure.
The txt2rdf process in a nutshell

- First find the important nouns, or “named entities”:

  **Example**
  
  Mr. Peter Moar found a small hoard of 5 polished stone knives on the same patch in May 1946.

- Then look for relationships between them:

  Mr. Peter Moar found a small hoard of 5 polished stone knives on the same patch in May 1946.
My txt2rdf pipeline

**Text documents**

- Pre-processing
  - Tokenise
  - Sentence and para split
  - POS tag

**Named Entity Recognition**

- Trained NER model
  - Multi-word tokens and features
  - List of NEs and classes

**Relation Extraction**

- Trained RE model
  - Set of NE pairs and features
  - List of relations and classes

**Graph of triples**

- RDF translation
  - Generate triples
  - Remove unwanted relations
  - Attach siteids

**Populating the Semantic Web**
txt2rdf
Dealing with EVENTs

Mr. Peter Moar found a small hoard of 5 polished stone knives on the same patch in May 1946.

- FIND event is n-ary relation:

  eventRel(eventAgent, eventAgentRole, eventDate, eventPatient, eventPlace)

  find123(“Mr. Peter Moar”,,“May 1946”,“polished stone knives”,“Hill of Shurton”)

Split event relation into RDF triples

:find123 :hasAgent :peterMoar .
:find123 :hasDate :may1946 .
:find123 :hasPatient :polishedStoneKnives .
:find123 :hasLocation :hillOfShurton .
Archaeological events

- Example was from RCAHMS data: http://canmore.rcahms.gov.uk/en/site/1102/details/hill+of+shurton/
- Extracting events allows structured searches
- Question: What do we do with NEs that are not in relations?

For more information
Available NLP systems

- OpenCalais: http://viewer.opencalais.com/
- Powerset: now part of Microsoft’s Bing
- AlchemyAPI: http://www.alchemyapi.com/api/demo.html
- Trained on lots of (English) text, mostly news articles
- Many general-purpose NLP tools available online, eg
  - nltk (http://www.nltk.org/)
  - Edinburgh LTG software (http://www.ltg.ed.ac.uk/software)
  - NaCTeM (http://nactem.ac.uk/)
Populating the Semantic Web
why and how?
text to RDF

Organising the Triples
schema design
open issues and problems

References
Do we have a data schema?

How do we populate the Semantic Web?

- create structured data
- convert unstructured and semi-structured
A schema for free text

- So far we’ve looked at instance data (ABox statements)
- Where is framework to slot instances into?
- Named Entity categories: PERSON, PLACE, DATE, etc
- NE categories become RDFS classes:

Mr. Peter Moar found a small hoard of 5 polished stone knives on the same patch in May 1946.

Entities are typed by their NE class
:peterMoar rdf:type :person .
:found rdf:type :event .
:may1946 rdf:type :date .
RDFS class hierarchy

- NE categories usually flat list, not hierarchical
- My set: ORG, PERSNAME, ROLE, SITETYPE, ARTEFACT, PLACE, SITENAME, ADDRESS, PERIOD, DATE, EVENT
- Contrast with structured data
- Hand-designed hierarchy (with rdfs:subClassOf):

```
http://www.ltg.ed.ac.uk/tether/
```

```
siteid  agent  loc  time  event  classn
  org  person  role  date  period  sitetype  objtype
  place  sitename  address
  grid

survey  excavation  find  visit  description  creation  alteration
```
Property labels

- Can we extract RDF predicate set automatically from text?
- Eg clustering commonly occurring verbs
- But – schema design only has to be done once
- Property set and class hierarchy are related
- My set: `hasEvent`, `hasAgent`, `hasAgentRole`, `hasPeriod`, `hasPatient`, `hasLocation`, `hasClassn`, `hasObject`, `partOf`
- Plus “standard” ones like `owl:sameAs`, `rdfs:seeAlso`, `skos:broader`
- Flat list (no `rdfs:subPropertyOf`)
Existing schemas and vocabularies

- Lots of vocabularies to choose from
  - http://esw.w3.org/topic/VocabularyMarket
- RDFS and OWL are fundamental
- SKOS – translate existing domain thesauri
  - http://www.w3.org/2004/02/skos/
- Reuse where possible; invent local scheme where not
- Incentive to reuse is strong
Issues around schema design

- **Schema granularity**
  - more categories to choose from $\Rightarrow$ NLP less accurate
  - so class hierarchy quite coarse and flat

- **Schema discovery**
  - tension between simplicity and query power
  - how will software agents “understand” your schema?
Issues around NLP accuracy

- Disambiguation of extracted NEs (Mr Peter Moar, P Moar, etc)
- NLP is not 100% accurate!
  - how many false statements can Semantic Web stand?
  - “the computer said it, so it must be true”
- Even given canonical URI :peterMoar for our Peter Moar...
- ... can we distinguish him from others with same name?
Grounding local data

- Grounding: tie unique canonical URIs to authoritative reference
- Is http://www.ltg.ed.ac.uk/tether/Loc/Place#edinburgh same as http://www.geonames.org/2650225/edinburgh.html?
- Grounding during NLP processing impractical
- Use owl:sameAs? – redundant triples, more complex queries
- Who maintains the authoritative lists of URIs?
- Specialist thesauri, eg archaeological classification terms like “Stone setting”
Grounding specialist terminology

Aim: ground place names, people, organisations, etc.
Linked Data
Organising the Triples: open issues and problems

Grounding turns data into Linked Data

Grounding local URIs against "authority" nodes is the next big challenge!
References to follow up

- **NLP software for Semantic Web applications**
  - AlchemyAPI: [http://www.alchemyapi.com/api/demo.html](http://www.alchemyapi.com/api/demo.html)

- **VocabularyMarket**: [http://esw.w3.org/topic/VocabularyMarket](http://esw.w3.org/topic/VocabularyMarket)

- **Case study**

- **General references for NLP**
  - *Speech and Language Processing*, Jurafsky and Martin ([http://www.cs.colorado.edu/~martin/slp.html](http://www.cs.colorado.edu/~martin/slp.html))