Datasets for Comparative Evaluation of RE in the Biomedical and News Domains

Ben Hachey, Claire Grover & Richard Tobin

bhachey@inf.ed.ac.uk

Overview

- We **re-factor** 2 relation extraction (RE) corpora to a single XML markup
  - BioInfer Biomedical RE corpus from Pyysalo et al. (2007)[1]
  - ACE News RE corpus from NIST[2] / LDC[3]
- We **re-annotate** ACE, automatically mapping nominal to named entities
- Result is a consistent and richly annotated **re-distribution**

Re-Factoring

A General IE Document Type

```
<!ELEMENT iedat (doc+)>          <!-- Iedat: Contains Doc(s)      -->
<!ELEMENT doc (text,markup)>     <!-- Doc: Contains Text, Markup -->

<!ELEMENT text (p)+>             <!-- Text: Contains paragraphs  -->
<!ELEMENT p (s|w)+>              <!-- P(aragraph): Contains Ss    -->
<!ELEMENT s (w+)>                <!-- S(entence): Contains Words -->
<!ELEMENT w (#PCDATA)>           <!-- W(ord): Contains Word Text -->

<!ELEMENT markup (nes,rels)>     <!-- Markup: Contains NEs, Rels -->
<!ELEMENT nes (ne*)>             <!-- Nes: Contains NE Mentions  -->
<!ELEMENT ne (textspan*)>        <!-- Ne: Contains NE Textspan   -->
<!ELEMENT textspan (#PCDATA)>    <!-- Textspan: Contains NE Text -->
<!ELEMENT rels (rel*)>           <!-- Rels: Contains Rel Ment'ns -->

<!ATTLIST doc id CDATA #IMPLIED>       <!-- Document ID        -->
<!ATTLIST s id CDATA #REQUIRED>        <!-- Sentence ID        -->
<!ATTLIST w id CDATA #REQUIRED>        <!-- Token ID           -->

<!ATTLIST ne id CDATA #REQUIRED>       <!-- NE Mention ID      -->
<!ATTLIST ne fr CDATA #REQUIRED>       <!-- NE Start Token ID  -->
<!ATTLIST ne to CDATA #REQUIRED>       <!-- NE End Token ID    -->
<!ATTLIST ne t CDATA #REQUIRED>        <!-- NE End Token ID    -->
<!ATTLIST ne st CDATA #IMPLIED>        <!-- NE Sub Type        -->

<!ATTLIST rel e1 CDATA #REQUIRED>      <!-- Rel NE 1 ID        -->
<!ATTLIST rel e2 CDATA #REQUIRED>      <!-- Rel NE 2 ID        -->
<!ATTLIST rel t CDATA #REQUIRED>       <!-- Rel Type           -->
<!ATTLIST rel st CDATA #IMPLIED>       <!-- Rel Sub Type       -->
```

Conversion

BioInfer Simple XML-to-XML transformation using XSLT

ACE Tokenise, then map from character to token offsets

1. Do sentence and word segmentation using LT TTT (Grover et al., 2000)
2. Convert from char to token standoff using LT XML2 (Grover et al., 2006)[4]

Example Document

```
<iedat>
...
<doc id='15'>
 <text>
  <p><s id='s11'><w id='w209'>Beta-catenin</w> <w id='w212'>is</w>
  <w id='w213'>also</w> <w id='w214'>found</w> <w id='w215'>in</w>
  <w id='w216'>these</w> <w id='w217'>structures</w><w id='w218'>.</w></s></p>
 </text>
 <markup>
  <nes>
   <ne id='e75' fr='w211' to='w211' t='Substance' st='Individual_protein'>
    <textspan>beta-catenin</textspan>
   </ne>
   <ne id='e77' fr='w217' to='w217' t='Source' st='Cell_component'>
    <textspan>structures</textspan>
   </ne>
  </nes>
  <rels>
   <rel id='r32' e1='e75' e2='e77' t='Causal' st='Change/Location'/>
  </rels>
 </markup>
</doc>
...
</iedat>
```

Linguistic Markup

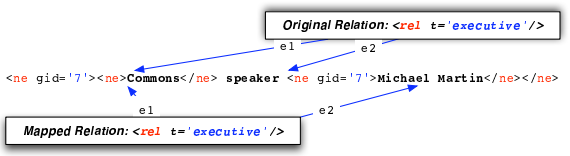
- We add part-of-speech, lemma, and detailed chunk markup to words
- In addition, we add Minipar dependency parse markup (Lin, 1998)

[1] http://mars.cs.utu.fi/BioInfer/
[2] http://www.nist.gov/speech/tests/ace/
[3] http://projects.ldc.upenn.edu/ace/
[4] http://www.ltg.ed.ac.uk/software/ltxml2

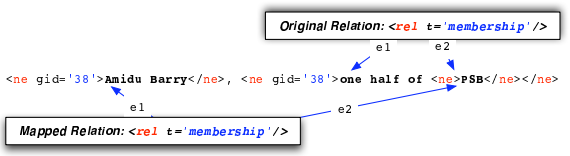
Re-Annotation

Mapping ACE

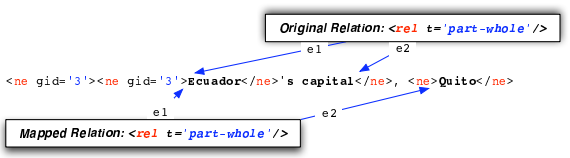
- ACE corpora are good resource for information extraction, but annotation is at deep linguistic (syntactic) level
  - Many complex, nested nominal entity references
  - Inconsistent with BioInfer and extrinsic tasks (e.g., summarisation)
- We automatically map many nominal entity references to named entity references using rules based on the detailed ACE annotation
- Total of 11 rules; Following 3 account for approx. 70% of mappings:



Mapping Rule 1: Prenominal ↦ Embedding Coreferent



Mapping Rule 2: Nominal ↦ Left Adjacent Coreferent



Mapping Rule 3: Nominal ↦ Right Adjacent Coreferent

Re-Distribution

- Re-factoring and re-annotation provide a consistent and richly annotated data set drawn from the ACE 2004, ACE 2005 and BioInfer corpora
- New resource suited to experiments with domain neutrality and adaptation
- We are preparing a distribution of the pipeline that creates our re-factored and re-annotated corpus directly from the original distributions
- To be available summer 2008 from `http://www.ltg.ed.ac.uk`

References

Grover, C., Matheson, C., Mikheev, A., and Moens, M. (2000). LT TTT—a flexible tokenisation tool. In *2nd LREC*, Athens, Greece.

Grover, C., Matthews, M., and Tobin, R. (2006). Tools to address the interdependence between tokenisation and standoff annotation. In *EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, Italy.

Lin, D. (1998). Dependency-based evaluation of minipar. In *LREC Workshop on Evaluation of Parsing Systems*, Granada, Spain.

Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., and Salakoski, T. (2007). BioInfer: a corpus for IE in the biomedical domain. *BMC Bioinformatics*, 8(50).