

The Alchemy of Annotation: When Biologists Disagree

Barry Haddow¹ Ewan Klein¹ Kirsty Lee¹
Elizabeth Fairley²

¹ Language Technology Group, School of Informatics,
University of Edinburgh and ² Cognition EU Ltd

20 March 2006

Collaborators

- ▶ Claire Grover
- ▶ Mike Matthews
- ▶ Leif Nielsen
- ▶ Malvina Nissim
- ▶ Stuart Roebuck
- ▶ Richard Tobin
- ▶ Xinglong Wang

Context

Creating an Annotated Corpus

Disagreements

Entities

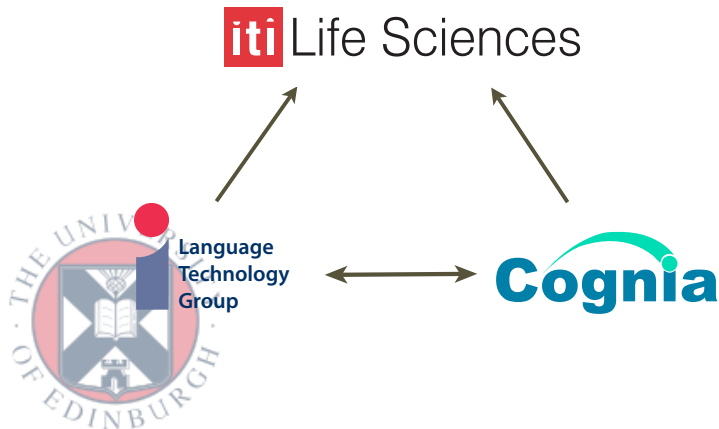
Relations

Conclusions

Project Information

- ▶ Text Mining Programme funded (3 yrs from Feb 2005) by ITI Life Sciences
 - ▶ General goal of ITI is to encourage market-driven research that can be commercialised in Scotland.
 - ▶ Programme is developing technologies for finding, retrieving and storing structured data from unstructured text.
 - ▶ Approach is intended to be generic, but current focus is on biological data in research papers.

Project Participants



Why Annotate?

- ▶ Annotated data plays two roles:
 - ▶ **Gold Standard** data for evaluation.
 - ▶ Training data.
- ▶ Annotated data should be informed by a domain ontology, as codified in Annotation Guidelines.

Annotation Decisions

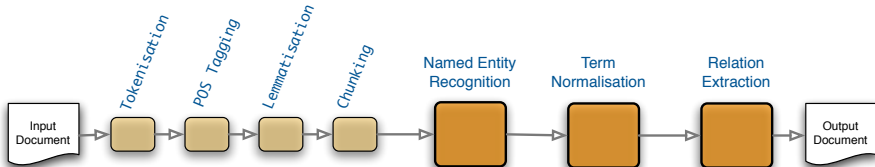
- ▶ Annotation decisions result from a variety of factors:
 1. the text,
 2. the guidelines,
 3. the annotator's background knowledge of the domain.
- ▶ There are also 'performance' issues:
 - ▶ tiredness, level of attention;
 - ▶ ergonomic aspects of the annotation tool.

What to Annotate?

- ▶ Our annotation is within context of standard Information Extraction paradigm
- ▶ Key markables:
 - Named Entities proteins, protein complexes, protein fragments
 - Relations protein (-complex) interactions
- ▶ NEs are normalized (relative to an ontology),
- ▶ and relations are stated between the normalized entities.

Text Processing Framework

- ▶ System uses LT-XML processing framework.
- ▶ XML mark-up is incrementally added at each stage.
- ▶ Incorporates both rule-based and statistical components.



Interactions vs. Associations

- ▶ Task was to identify protein-protein interactions which are
 - ▶ **direct** — proteins are in physical contact.
- ▶ These were categorized with the label **PPI**.
- ▶ Relations which are were judged to be **indirect**, or where the annotator was unsure about the direct/indirect status were labeled 'protein-protein associations' (**PPA**).

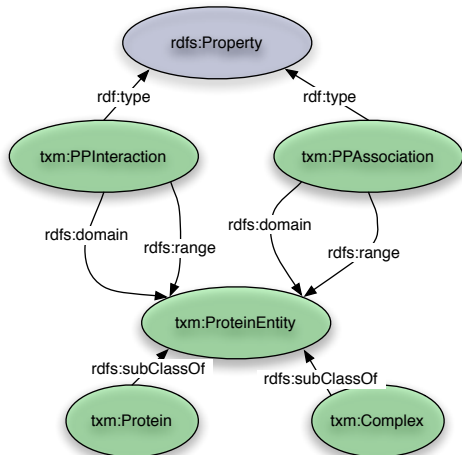
Example Protein-Protein Interaction

Direct

To examine interactions between $ER\beta$ and N-CoR in mammalian cells we performed two-hybrid assays using a GAL4 DBD/N-CoR C-terminus fusion protein as bait and a VP16- $ER\beta$ LBD fusion as the prey. Fig. 2 shows that the **$ER\beta$ -LBD bound N-CoR** in the presence of agonists and phytoestrogens, but not SERMs.

- ▶ This example includes information about experimental support, which is typically not present in abstracts.

Simplified RDFS Ontology



22 Carat Gold?

- ▶ How good is a given corpus of annotated data?
 1. Is it a **representative** sample?
 2. Is there **enough** data for training?
 3. Is it an **accurate** reflection of domain knowledge?
 4. Is it **internally consistent**?
- ▶ In this talk, mainly concerned with (3) and (4).
- ▶ Earlier work on BioCreative vs. BioNLP indicates that consistency of annotation significantly affects quality of learned model.

Pilot

Full text	6
Abstracts	18
Total	24

Table: Pilot — quadruply annotated documents

- ▶ Goals of pilot annotation exercise:
 - ▶ train annotators
 - ▶ refine and debug annotation tool
 - ▶ refine and debug annotation guidelines
- ▶ Four biologists annotated whole set of documents.
- ▶ Fifth (senior) biologist carried out reconciliation of differences.

Data: Quantity

- ▶ Sample of 7,932 documents from PubMed and PubMed Central selected by keyword.
- ▶ Manually reviewed for relevance.

	All	Dble
Full text	151	10
Abstracts	750	36
Total	901	46

Table: No. of documents annotated

Data: Distribution

protein	44,306
fragment	2,438
complex	2,362
exp. protein	1,596
fusion	943
Total	50,049

Table: Protein Entity Subtypes

PPI	1,566
PPA	4,173
Total	5,739

Table: Relations

Calculating IAA, 1

- ▶ Annotation consistency: Inter-Annotator Agreement (IAA).
- ▶ For some set D of doubly annotated documents, let M_1 , M_2 be the number of 'markables' in D marked-up by annotators 1 and 2 respectively.
- ▶ Let C be the total number of times where both annotators agree on a markable.
- ▶ Then F-score is defined to be $F = \frac{2 \times C \times 100}{M_1 + M_2}$
- ▶ So $F = 100$ iff there is perfect agreement, and $F = 0$ iff there is no agreement.

Calculating IAA, 2

Annotator A	Annotator B
m_1	
m_2	m_2
	m_3
m_4	m_4

- ▶ For relations, we only considered cases where
 - ▶ the proteins in the putative relation were in the same sentence, and
 - ▶ both entities had in fact been labeled as proteins by both annotators.

IAA Figures

- ▶ Same four biologists as in Pilot carried out annotation.
- ▶ IAA figures arrived at by examining all 46 doubly-annotated documents then averaging.

Named Entity	87.61
Term Normalization	69.55
Relation Extraction	52.22

Table: F-scores for Inter-Annotator Agreement

Analysing Sources of Disagreement

- ▶ In order to inform future efforts on annotation, we need to understand current sources of disagreement between annotators.
- ▶ Work is in progress to examine doubly annotated data, and to categorize the disagreements.
- ▶ Results to date are **preliminary**, and require further scrutiny by biologists.

Disagreeing about Entities

Type of Disagreement	Count	Category
protein / complex	7	bio
protein / \emptyset [gene]	12	bio
protein / \emptyset [other]	59	bio
complex / \emptyset [other]	32	bio
fragment / \emptyset [other]	15	bio
protein as modifier	6	ling
coordinates	15	ling
entity boundary	5	ling

Table: Top 80% sources of disagreement on sample of 175

Only 15% of the sample seem to involve linguistic issues rather than biological ones.

Relations: Levels of Disagreement

Match Condition	No. relations annotated	No. agreements	F score
strict	406	35	17.24%
relaxed	406	106	52.22%
PPI only	151	17	22.52%
PPA only	255	18	14.12%

Table: Inter-Annotator Agreement on Relations

Match Conditions

strict: both annotators agree on presence of a relation **and** on its type (PPI vs. PPA).

relaxed: PPI and PPA are collapsed into one type.

Relations: Types of Disagreement

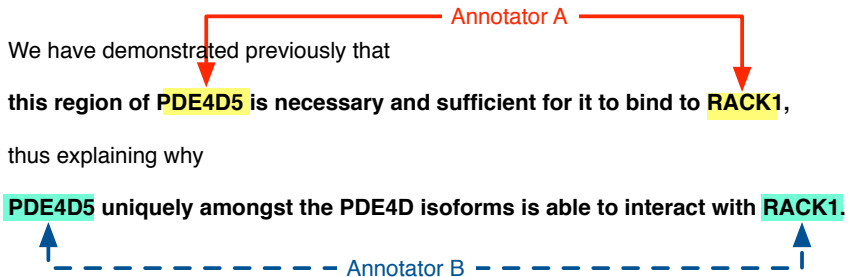
Type	Count	Description
interaction / \emptyset [error]	41	Disagreement is due to annotation error
interaction / \emptyset [ambig]	33	Both annotations plausible
ppi-ppa	31	Disagreement over relation type

Table: Sources of disagreement: relations

Type	Count	Description
pathway	10	Participant is pathway, not protein
family	1	Participant is family, not protein
other	4	Participant is other non-protein
abbrev	8	Protein synonyms not consistently marked

Table: Sources of disagreement: relation participants

Distributed Information



Worldviews

- ▶ Biologist's orientation to the text:
 - ▶ Has a putative PPI been experimentally proven at this point in the paper?
 - ▶ Is the evidence convincing?
- ▶ Linguists's orientation to the text:
 - ▶ What formal means does the text use to describe PPIs?
 - ▶ Are there consistent and reliable textual clues for what we want to learn?

Negated Sentences

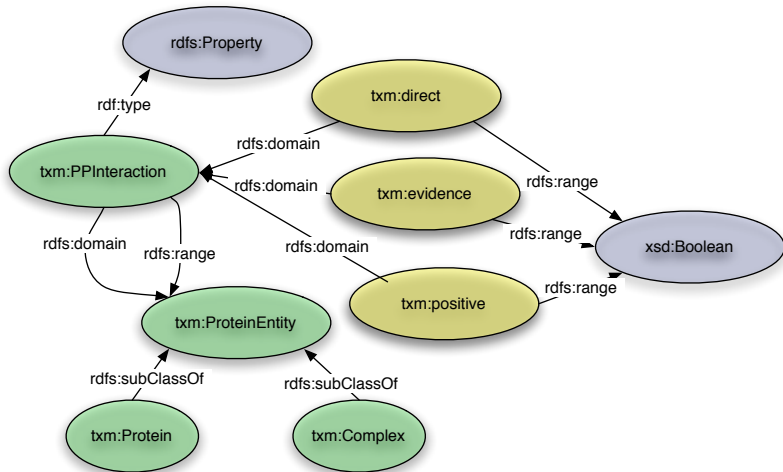
Our preliminary work suggests that CREB-H does not interact directly with hepatitis C virus core protein. (pmc_1072803_ft)

EAS-3 is not phosphorylated by pSAPK or pp38.
(pmc_1180429_ft)

Key Conclusions

- ▶ Although there were good number of clear cases of PPI, distinguishing between PPIs and PPAs seemed too unreliable (34% of disagreements).
- ▶ Proposed solution is to refactor and extend the ontology:
 - ▶ basic property of **interaction**;
 - ▶ additional 'facets'
 - ▶ **direct** vs. **indirect**
 - ▶ **proven** vs. **referenced**
 - ▶ **positive** vs. **negative**

Revised RDFS Ontology



Summary

- ▶ IAA gives a good indication of internal consistency of annotated data.
- ▶ Higher IAA also leads to more confidence in accuracy.
- ▶ IAA is standardly taken as a ceiling for supervised learning approaches.
- ▶ In order to make progress on relation extraction, high quality training and test materials play a vital role.
- ▶ Low IAA for relation annotation suggests task is intrinsically difficult.
- ▶ Nevertheless, analysis of disagreement suggests better agreement can be achieved.