# Biomedical Event Annotation with CRFs and Precision Grammars

Andrew MacKinlay, David Martinez, Timothy Baldwin

# Motivation and Architecture

NICTA

- Our motivation: deep linguistic processing for detection of speculation and negation
- Architecture:
  - Task 1:
    - Trigger word detection: CRF and Lookup systems
    - Event-theme construction (hand-crafted rules)
  - Task 3:
    - Deep parsing for semantic representation
    - Classification of events using Maximum Entropy

# Trigger word detection with CRFs

- Conditional probability distribution over label sequences given a particular observation sequence
- CRF++ toolkit (Sha and Pereira, 2003)
- Tested features: word-form, lemma, POS, chunking marks, protein NER, grammatical dependencies (from Bikel parser and GDep)
- JULIE-Lab sentence splitter and Genia Tagger for pre-process
- Window sizes: $\pm 3$ and $\pm 4$

# Trigger word detection with CRFs

- Best results (training data): Precision $\sim$ 66%, Recall $\sim$ 30%
- All features help except for grammatical dependencies
- $\pm$3 window size

# Trigger word detection with dictionary look-up

- Decision list for each trigger string found in training data
  - Simply assign highest frequency class
- Frequency cut-off
- We can reach high recall ($\sim$ 77%) but at the cost of precision ($\sim$ 13%)
- Best f-score $\sim$ 36% ($\sim$ 50% recall)

# Trigger word detection: Combination

- Add all trigger words identified by CRF and look-up
- Two approaches:
  - Optimise per class (Optim)
  - Always preference to CRF (All)

# Event-theme construction

- Approach: assign closest events/proteins as themes (without crossing sentence boundaries)
- Basic events:
  - Single closest protein
- Binding events:
  - Closest proteins
  - Parameters: maximum distance and number of themes
- Regulation events
  - Single closest protein or event (give precedence to events)
  - Parameters: maximum distance and detect/ignore CAUSE

# Task 1 Results

| System | Rec. | Prec. | FSc. |
|---|---|---|---|
| **Combined (Optim.)** | **17.44** | **39.99** | **24.29** |
| Combined (All) | 24.36 | 30.87 | 27.23 |
| CRF | 12.23 | 62.24 | 20.44 |
| CRF (+ synt feats) | 12.01 | 61.91 | 20.11 |
| Look-Up | 22.88 | 29.67 | 25.84 |
| Look-Up (freq $>=$ 20) | 23.26 | 26.74 | 24.88 |
| Look-Up (freq $>=$ 30) | 21.37 | 30.50 | 25.13 |

Table: Task 1 results with approximate span matching, recursive evaluation (our final submission is in bold)

# Negation/Speculation detection

- English Resource Grammar (ERG): high-precision grammar in the HPSG framework
- GENIA tagger to deal with named entities
- 72% of training sentences parsed

# Feature extraction

- Semantic formalism: Robust Minimal Recursion Semantics
- Elementary Predicates (EP): Predicates with their arguments
- Relationships between trigger EP and lexical cues
  - Outscoping and shared-argument

# Features for negation identification

- Pre-identify word lists:
  - Conjunctions: *_not_c*, *_but+not_c*, *_nor_c*
  - Other markers: *_only_a*, *_never_a*, *_not+as+yet_a*, *_not+as+yet_a*, *_unable_a*, *neg_rel*
- Negative-outscope feature: when negative EP outscopes trigger-EP
  - E.g. "...product was **not (NEG-EP)** able to **bind (TRIG-EP)** DNA and..."
    - NegOutscope neg_rel = 1
    - NegOutscope not = 1

# Features for negation identification

- ...product was **not** able to **bind** DNA and was recovered in cytoplasmic cellular extracts...
- ERG analysis
  - **l8**: neg_rel⟨692 : 695⟩(e9, ARG1: **h10**)
  - **l11**: _able_a_1⟨696 : 700⟩(e12, ARG1: x6, ARG2: **h13**)
  - **l14**: _bind_v_to⟨704 : 708⟩(e17, ARG1: x6, ARG2: x16, ARG3: u15)
  - h10 qeq **l11**, h13 qeq **l14**
- Thus l8 immediately outscopes l11, and l11 immediately outscopes l14

# Features for negation identification

- Negative conjunction: when trigger-EP is the argument (ARG0) of a negative conjuction EP
  - E.g. “...**but not (NEG-EP) binding (TRIG-EP)** DNA...”
- When trigger-EP is the argument (ARG0) of a negatively-outscoped EP
  - E.g. “...the **product (TRIG-EP)** was **never (NEG-EP)** considered...”

# Features for speculation identification

- Pre-identify word lists:
  - Speculation verb short list: _*investigate*, _*study*, _*examine*, _*test*, _*evaluate*, _*observe*}
  - Extended list: adding WordNet sisters
- SpecVOBJ: when verb part of “speculative-verbs” set, and object is a trigger word
  - E.g. “IkappaBalpha phosphorylation and **degradation (TRIG-EP)** was **analyzed (SPEC-EP)**”
    - SpecVObj2+WN-seed:examine = 1
    - SpecVObj2+wn-sister:_analyze_v_1(examine) = 1
    - SpecVObj2+wn-gen = 1

# More features

- Speculation:
  - Modal verb outscopes trigger
  - ARG0 of trigger-EP occurs as argument of the word 'analysis'
- General features:
  - E.g. (Modifier adjective) "...Fas **upregulation (TRIG-EP)** is **central (ADJ-EP)** to the preservation..."
  - 'ModAdj:_central_a_1' = 1
  - Trigger name, trigger POS, etc.

# Negation/Speculation Classifiers

- Maximum Entropy classifier (Maxent Toolkit)
- Different feature combinations
- Baseline: bag of words
- Development phase:
  - Goldstandard events
  - 10-fold cross-validation
- Test phase:
  - Trained over goldstandard event extraction
  - Output of task-1 classifier as source of trigger words

# Development results: Speculation

| Feats. | Rec. | Prec. | FSc. |
|---|---|---|---|
| BOW | 22.1 | 47.7 | 30.2 |
| Spec. + BOW | 23.2 | 57.9 | 33.1 |

- Very low performance over automatic classification
- Linguistic features better than BOW
- Combination of features works best

# Development results: Negation

| Feats. | Rec. | Prec. | FSc. |
| --- | --- | --- | --- |
| BOW | 15.0 | 30.2 | 20.0 |
| Neg. + BOW | 24.3 | 68.4 | 35.9 |

- Bigger improvement over BOW

# Official results for Task 3

| TEAM | gold (match) | answer (match) | recall | prec. | fscore |
|---|---|---|---|---|---|
| ConcordU | 3617 ( 1182) | 1943 ( 1182) | 32.68 | 60.83 | 42.52 |
| VIBGhent | 3617 ( 1105) | 2227 ( 1104) | 30.55 | 49.57 | 37.80 |
| ASU+HU+BU | 3617 ( 710) | 1185 ( 710) | 19.63 | 59.92 | 29.57 |
| **NICTA** | 3617 ( 577) | 1450 ( 575) | 15.95 | 39.66 | **22.75** |
| USzeged | 3617 ( 722) | 3113 ( 722) | 19.96 | 23.19 | 21.46 |
| CCP-BTMG | 3617 ( 446) | 777 ( 446) | 12.33 | 57.40 | 20.30 |

# Lessons learned

- Keyword detection suffers from data sparseness
- Rules for event construction are too naive
- Deep parsing better than lexical baseline, but there are coverage problems
- Combined approach (detect triggers and themes together) to be explored for task 1

# Biomedical Event Annotation with CRFs and Precision Grammars

Andrew MacKinlay, David Martinez, Timothy Baldwin

Australian Government
Department of Broadband, Communications and the Digital Economy
Australian Research Council

NICTA Members
ANU THE AUSTRALIAN NATIONAL UNIVERSITY
UNSW THE UNIVERSITY OF NEW SOUTH WALES
Department of State and Regional Development

Victoria The Place To Be
THE UNIVERSITY OF MELBOURNE
The University of Sydney
Queensland Government
Griffith UNIVERSITY
QUT
THE UNIVERSITY OF QUEENSLAND AUSTRALIA
NICTA Partners