

Event Extraction on PubMed Scale

Filip Ginter

University of Turku

Finland

`http://bionlp.utu.fi`

Acknowledgments

This talk is based on joint work with:

- **Jari Björne**, University of Turku
- **Sampo Pyysalo**, University of Tokyo

Also many thanks to BioTM organizers!

Event extraction

- *Events* are a common information extraction target in the bio-domain
- Aim to capture the knowledge at a reasonable detail
- Event extraction popularized by the *BioNLP'09 Shared Task on Event Extraction*
- ...and Sampo's talk yesterday
- In this talk: events as defined by the Shared Task (based on the GENIA Event corpus)

Events (in context of the Shared Task)

- Events as defined by the BioNLP'09 Shared Task on Event Extraction (Tasks 1–3):
- **Typed:** Gene Expression, Transcription, Localization, Binding, Positive regulation, . . . (9 types)
- Any number of **arguments** with associated **roles**
 - *Cause* and *Theme* in the basic form (Task 1)
 - *AtLoc*, *ToLoc*, *Site*, *CSite* in enriched form (Task 2)
- **Recursively-nested** — events can be arguments of other events
- **Negated and/or speculated** (Task 3)

Event type	Example
Gene expression Transcription	<u>5-LOX</u> is <i>expressed</i> in leukocytes promoter associated with <u>IL-4 gene transcription</u>
Localization	phosphorylation and nuclear <i>translocation</i> of <u>STAT6</u>
Protein catabolism	I kappa B-alpha <i>proteolysis</i> by phosphorylation.
Phosphorylation	<u>BCL-2</u> was <i>phosphorylated</i> at the G(2)/M phase
Binding	<u>Bcl-w</u> <i>forms complexes</i> with <u>Bax</u> and <u>Bak</u>
Regulation	<u>c-Met expression</u> is <i>regulated</i> by <u>Mitf</u>
Positive regulation	<u>IL-12</u> <i>induced</i> <u>STAT4 binding</u>
Negative regulation	<u>DN-Rac</u> <i>suppressed</i> <u>NFAT activation</u>

BioNLP'09 Shared Task

The Shared Task was a carefully controlled experiment:

- Gold-standard named entities
- Relatively narrow domain defined by the MeSH terms *human*, *blood cells*, and *transcription factors*
- Focus lies on event extraction, as opposed to events themselves
- Participants concentrate on optimizing the primary performance measure on the test set of 260 PubMed abstracts
- Training, development, and evaluation data sampled from the same corpus
- The goal of the Shared Task: to advance information extraction in the bio domain
- The result of the Shared Task: *NLP methods and software*

BioNLP'09 Shared Task (cont.)

- The Shared Task systems are *not* the goal as far as *BioNLP users*, as opposed to BioNLP researchers, are concerned
- The target audience is interested in the output of the systems
- Most of the Shared Tasks systems have not been used outside of the Shared Task
- Even though at least some of the systems are available, they might be tedious to run
- Put the Shared Task systems to use!

Going PubMed-scale

- Only relatively small sets of events exist so far: 13,600 manually annotated events in the entire Shared Task data from 1,210 PubMed abstracts
- Shared Task system output on the 260 abstracts in the test set
- Too small to be useful outside of BioNLP research
- For many of the speculated applications of event extraction (and biomedical IE as such), one would hope for a much larger scale
- Why not go all the way to the 11M abstracts in PubMed?
- ...and throw in the titles of all 17.8M PubMed citations for good measure...

Going PubMed-scale (cont.)

- Can this be done?
- Scalability was not a requirement in the Shared Task
- The titles + abstracts total some 200M sentences of text
- What does it take to scale event extraction up $40,000\times$?
- What are the problems involved and what output do we get?

The pipeline

- We process the PubMed titles and abstracts with the following pipeline:
- NER: The BANNER system
- Sentence splitter: GENIA group's ML-based splitter + rule-based post-processing for systematic issues
- Tokenizer: Parser's own
- Parser: McClosky-Charniak-Johnson + Stanford scheme conversion
- Event extraction: The Turku system

STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .

A

named entity recognition

Protein

STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .

Protein

Protein

B

parsing

<nn> <nsubj> <dojb>
 <appos> <aux> <dojb> <conj_and>
 NNP NNP NN MD VB NNP CC NNP
 STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .

C

event detection

<Theme> <Theme> Regulation Cause>
 <Site> <Theme> Cause>
 Protein Entity Phosphorylation Regulation Protein Protein
 STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .

D

speculation and negation detection

Spec
 <Theme> Regulation Cause>
 <Site> <Theme> Cause>
 Protein Entity Phosphorylation Spec Regulation Protein Protein
 STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .

The pipeline (cont.)

- When we started, all components of the pipeline had the best, or near best, reported performance on their task:
- BANNER: near-best on BioCreative II dataset, best in comparison by Kabiljo et al. [1]
- McClosky-Charniak-Johnson: best so far reported parsing performance on GENIA Treebank
- Turku system: best in the Shared Task (journal extension: best on all three sub-tasks)

By these results, we can expect the output of the pipeline to be about the best we are currently capable of using publicly available software.

Event extraction from PubMed

Event extraction outcome on all titles and abstracts from PubMed:

- NER: **36.5M NE mentions** in **5.4M citations**
- Parser: **20M syntactically parsed sentences** (those with at least one NE mention)
- Event extraction: **19.2M event occurrences** (4.5M unique) of which 2.1M occurrences (1.6M unique) recursively contain at least two different NEs

Time requirements

Time requirements in CPU hours:

NER	1,800h	20%
Parsing	5,000h	60%
Event Extraction	1,500h	20%
Total	8,300h	

- 8,300h equals approx. one CPU year (346 CPU days)
- Only few weeks when parallelized on a cluster system
- These are ballpark figures, of course
- Real time requirements a little higher due to pipeline restarts, etc.

Technical issues

- Not an easy ride at times
- Rare issues get greatly magnified at this scale
- Many standard tools start failing in unpredictable ways under the load
- Workarounds needed at all levels:
 - Wrap almost anything with a layer of code to restart it whenever it crashes, or kills it whenever it runs for too long
 - Copious cross-checks across the pipeline
- We hope we got it right. . . :)
- . . . but let's get back to the events. . .

Output quality

- The pipeline components were previously evaluated separately on a relatively uniform, in-domain data
- What happens when we take the components out of their domain is a big unknown
- Easy to estimate precision by random sample and manual inspection
- The following numbers established by drawing 100 random cases (events/entities) and inspecting manually:
 - NER: 87% precision (compare to 87% reported on BioCreative II data)
 - Event extraction: 64% precision (compare to 58% in the Shared Task)
 - Often difficult to decide whether an event is correct or not (borderline cases, personal preferences vs. GENIA annotator guidelines)

Output quality (cont.)

- Task 2 arguments (site and location): 53% precision (compare to 58% on the Shared Task data)
- Negation (among otherwise correct events predicted as negated): 82% correctly predicted as negated
- Speculation (among otherwise correct events predicted as speculated): 88% correctly predicted as speculated
- Difficult to estimate recall
- The scale of the data compensates for whatever the recall is to some extent
- With $\sim 60\%$ precision of the core events, the dataset still contains millions of correct events

Normalization — arguments

- For the purpose of any event occurrence statistics, we need to establish event equality
- Two events are equal when:
 - Types are equal
 - All Task 1 arguments (*cause* and *theme*) are recursively equal
 - Arguments are compared in a canonical order

regulation(Cause:A, Theme:binding(Theme:B, Theme:C))
equals to
regulation(Theme:binding(Theme:C, Theme:B), Cause:A)

Normalization — entities

- Prefixes and suffixes: *p53* vs. *human p53 protein*
- Prefixes and suffixes with a predictable “near-equality” relation between the underlying gene or protein name and the entire automatically tagged entity mention are extracted from the dataset of static relations by Pyysalo et al. [2]
- We consider only those prefixes and suffixes whose removal results in a more frequent named entity (as measured in the PubMed data, not Pyysalo et al. data):
- *p53* tagged 117,000 times while *p53 protein* only 12,000 times → *protein* is a candidate suffix for removal in this case
- *capsid* tagged 7 times while *capsid protein* 2,000 times → *protein* is not a candidate for removal in this case
- Produces a replacement table *named entity* → *underlying gene/protein name*
- Mapping applied to 12% unique strings (6% tagged occurrences)

Normalization — entities (cont.)

term	underlying G/P name
p53 protein	p53
p53 gene	p53
human serum albumin	serum albumin
wild-type p53	p53
c-fos mRNA	c-fos
endothelial NO synthase	NO synthase
MHC cl. II molecules	MHC cl. II
human insulin	insulin
HIV-1 rev.transcriptase	rev.transcriptase
hepatic lipase	lipase
...	

Normalization — entities (cont.)

- 1. apply the aforementioned mapping
- 2. remove all non-alphanumeric characters and lowercase
- Addresses minor typographic variants: *TNF1*, *TNF-1*, *TNF 1*, *tnf1*, ...
- Two entities are considered equal when their normalized forms are equal
- Only an approximate solution to the quite complex problem
- Will suffice for first glance on the data, but needs an improvement later

Event recurrence

- **All events**

19.2M occurrences of 4.5M unique events → 4.2× repetition

- **Events with at least two different named entities**

2M instances of 1.6M unique events → 1.3× repetition

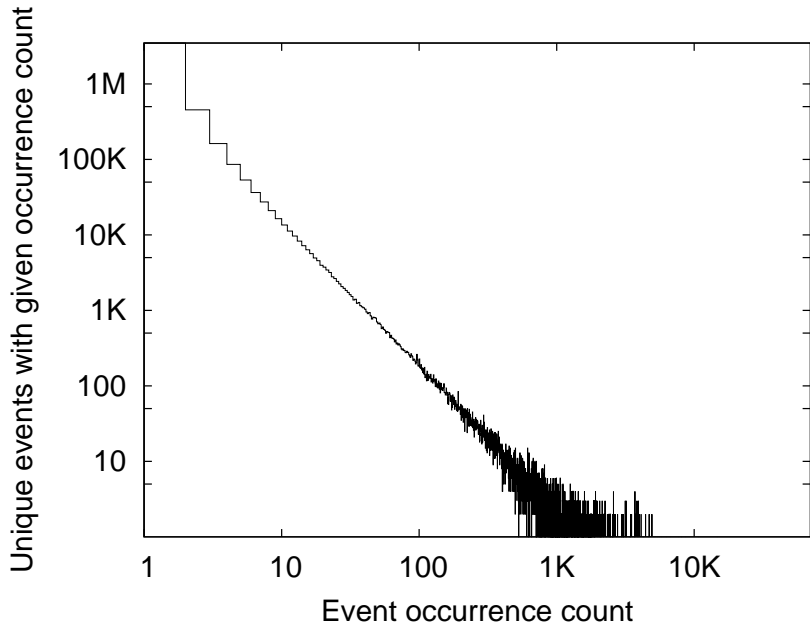
- Long-tailed distribution

- Most common event: *expression(Theme:insulin)* 59,821×

- Most common 2+ entity event: *positive-regulation(Cause:GnRG,Theme:localization(Theme:LH))*
699×

- Gonadotropin-releasing hormone, Luteinizing hormone — important in human reproduction

- 3.8M events are singleton



Event types

Event type counts among the 4.5M unique events extracted:

Positive regulation	1.2M
Binding	1M
Gene expression	700K
Negative regulation	670K
Regulation	560K
Localization	147K
Transcription	79K
Phosphorylation	49K
Protein catabolism	39K

Citation topics

- Can we say something about the topics of event-containing citations?
- Topics \leftrightarrow MeSH terms
- For each event type, take five MeSH terms with the highest point-wise mutual information
- $\log \frac{P(\text{event type, MeSH term})}{P(\text{event type})P(\text{MeSH term})}$
- Prior to the calculation, the original MeSH term set of each citation is expanded to all hypernyms

Event type	Five most related MeSH descriptors
Gene expression	Gene Expression Regulation; RNA; Gene Expression; Cytokines; Immunohistochemistry
Positive regulation	Intracellular Signaling Peptides and Proteins; Phosphotransferases; Transcription Factors; Cytokines; Gene Expression Regulation
Negative regulation	Molecular Mechanisms of Pharmacological Action; Intracellular Signaling Peptides and Proteins; Therapeutic Uses; Phosphotransferases; Tumor Cells, Cultured
...	

Citation topics (cont.)

- How well do these MeSH terms delimit citations with events?
- 37% of PubMed citations are indexed by at least one of the MeSH terms or its hyponym (*MeSH-relevant citation*)
- These contain 93% of all extracted events
- The MeSH terms distinguish very well citations from which the system extracts events
- But do they distinguish citations from which the system extracts *correct* events?
- Are MeSH-relevant citations more likely to contain correctly extracted events?

Citation topics (cont.)

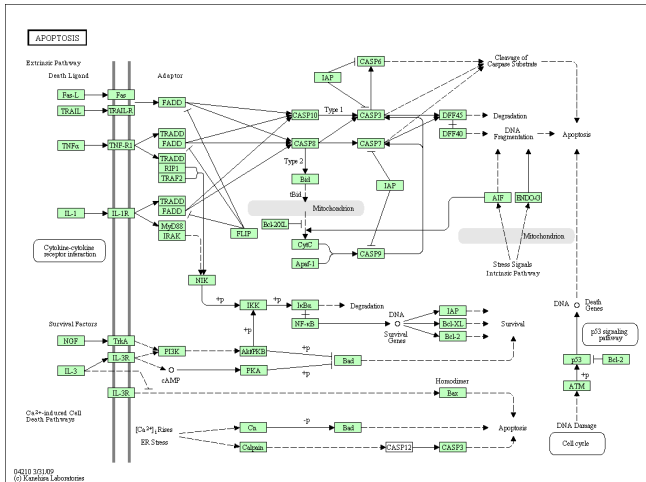
Using the 100 random events / entities used previously to estimate the precision:

	Citation	TP	FP	precision
Named entities	MeSH-relevant	66	5	93.0%
	MeSH-irrelevant	21	8	72.4%
Events	MeSH-relevant	58	29	66.7%
	MeSH-irrelevant	4	9	30.8%

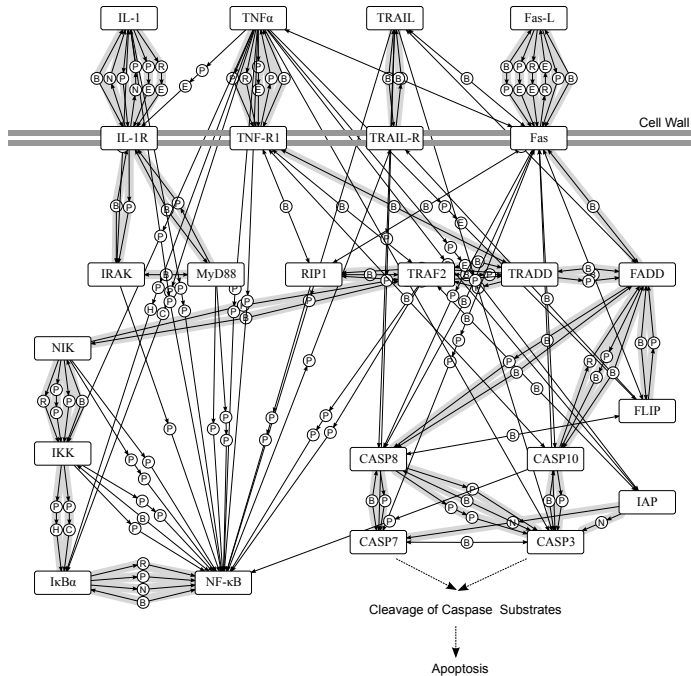
- 21 percentage points difference in precision of named entities (93-72)
- 36 percentage points difference in precision of events (67-31)
- Both differences statistically significant (two-tailed Fisher's test)
- Overall precision only grows by 2.7 percentage points (since so few events are extracted from MeSH-irrelevant citations)

Diving into the data

- We have so far looked at overall statistics of the data
- Let us “zoom in” to part of the apoptosis pathway as a point in case
- Difficult to fit the extraction output into one slide, so:
 - For each gene/protein we show:
 - 4 most common events in the extraction output and
 - 4 most common events present the actual apoptosis pathway in KEGG
 - (there can be overlap between these two sets)
 - Pruned down to NE pairs for clarity
- Protein synonyms manually resolved in the apoptosis pathway analysis

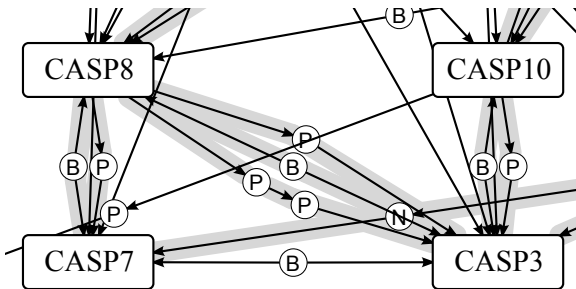
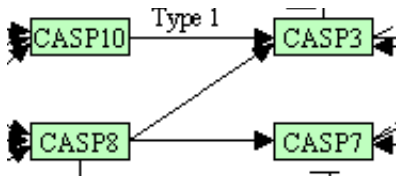


<http://www.genome.jp/kegg/pathway/hsa/hsa04210.html>



Apoptosis pathway

- The most common event was the positive regulation of $NF\kappa B$ by $TNF\alpha$
- The caspase part of the pathway is correctly recovered, including $CASP3$ — $CASP7$ which is not on the KEGG pathway figure



Apoptosis pathway (cont.)

IKK—I κ B α —NF κ B

- *IKK* phosphorylates *I κ B α* , launching its breakdown
 - 15× positive-regulation(C:*IKK*,T:catabolism(T:*I κ B α*))
 - 14× positive-regulation(C:*IKK*,T:phosphorylation(T:*I κ B α*))
- *I κ B α* negatively regulates *NF κ B* by binding and trapping it
 - 124× binding(T:*I κ B α* ,T:*NF κ B*)
 - 21× negative-regulation(C:*I κ B α* ,T:*NF κ B*)
 - 17× positive-regulation(C:*I κ B α* ,T:*NF κ B*)
 - 16× regulation(C:*I κ B α* ,T:*NF κ B*)
- The resulting positive regulation of *NF κ B* by *IKK* is extracted as well
 - 43× positive-regulation(C:*I κ B α* ,T:*NF κ B*)
 - 28×
positive-regulation(C:*I κ B α* ,T:positive-regulation(T:*NF κ B*))
 - 29× binding(T:*I κ B α* ,T:*NF κ B*)

- We did this so you don't have to :)
- Everything will be downloadable
- Two types of resources:
 - The extracted events themselves — the output of the system
 - The output of the event extraction pipeline stages

- Resource for those who are interested in the events themselves
- All .a1 and .a2.t123 stand-off files (Shared Task format) tarred in batches of 30,000
- Classifier prediction strengths for the extracted events
- A single XML with all events highly aggregated to achieve a manageable size and easy processing
- An SQL database under consideration

```
<ev c="309">
  <e r="C,T" t="Positive_regulation">
    <n>IL-4</n>
    <e r="T" t="Gene_expression">
      <n>IgE</n>
    </e>
  </e>
</ev>
```

Other data

- Resources for those who are interested in event extraction system development
- All identified named entities (the .a1 files mentioned previously)
- The output of the McClosky-Charniak-Johnson parser in the PENN bracketed format for nearly all (>99.9%) sentences with at least one named entity (20M sentences)
- Conversion of the PENN trees into collapsed, cc-processed Stanford Dependencies

What next? — Name normalization

- Current normalization not sufficient → misses name synonyms
- Point in case: the *FLIP* protein in the apoptosis pathway
- Currently no connection to existing bio-databases
- The dataset is open for anyone to use and we'd be happy to collaborate on the normalization

FLIP is a protein in the apoptosis pathway, known under all these names:

Protein: CASP8 and FADD-like apoptosis regulator, Cellular FLICE-like inhibitory protein, c-FLIP, Caspase-eight-related protein, Casper, Caspase-like apoptosis regulatory protein, CLARP, MACH-related inducer of toxicity, MRIT, Caspase homolog, CASH, Inhibitor of FLICE, I-FLICE, FADD-like antiapoptotic molecule 1, FLAME-1, Usurpin

Gene: CFLAR, CASH, CASP8AP1, CLARP, MRIT

What next? — Visualization

- **Visualization**
- We deal with 4.5M unique events
- Typed, multi-argument, recursively nested
- With prediction confidences and occurrence counts
- Linking back to PubMed where they were extracted from
- Possibly also linking to databases like UniProt
- How to present all this information in a useful way?
- Who are the possible users of this dataset and for what purpose?

What next? — Event network refinement

- Can the events be refined in context of other extracted events?
- Particularly taking into account the size of the dataset
- Identifying biologically unlikely event combinations
- Possible connection with known interaction databases (this would require normalization)
- Could, e.g., seed an iterative refinement process with certainly correct events

What next? — More data!

- The pipeline is set up and primed
- Full text articles are an obvious target
- Must beware of the copyrights, though

<http://bionlp.utu.fi/biotm.html>

- Björne, Ginter, Pyysalo, Tsujii, Salakoski: Complex Event Extraction at PubMed Scale. Proceedings of ISMB'10 (To appear.)
- One paper currently under review

Thank You!



R. Kabiljo, A. Clegg, and A. Shepherd.

A realistic assessment of methods for extracting gene/protein interactions from free text.

BMC Bioinformatics, 10(1):233, 2009.



S. Pyysalo, T. Ohta, J.-D. Kim, and J. Tsujii.

Static relations: a piece in the biomedical information extraction puzzle.

In *Proceedings of the BioNLP 2009 Workshop*, pages 1–9, Boulder, Colorado, June 2009. Association for Computational Linguistics.