Scaling up Biomedical Event Extraction to the Entire PubMed

Jari Björne[*,1,2] **Filip Ginter,**[*,1] **Sampo Pyysalo,**[*,3] **Jun'ichi Tsujii,**[3,4] **Tapio Salakoski**[1,2]
[1]Department of Information Technology, University of Turku, Turku, Finland
[2]Turku Centre for Computer Science (TUCS), Turku, Finland
[3]Department of Computer Science, University of Tokyo, Tokyo, Japan
[4]National Centre for Text Mining, University of Manchester, Manchester, UK
jari.bjorne@utu.fi,ginter@cs.utu.fi,smp@is.s.u-tokyo.ac.jp
tsujii@is.s.u-tokyo.ac.jp,tapio.salakoski@it.utu.fi

Abstract

We present the first full-scale event extraction experiment covering the titles and abstracts of all PubMed citations. Extraction is performed using a pipeline composed of state-of-the-art methods: the BANNER named entity recognizer, the McClosky-Charniak domain-adapted parser, and the Turku Event Extraction System. We analyze the statistical properties of the resulting dataset and present evaluations of the core event extraction as well as negation and speculation detection components of the system. Further, we study in detail the set of extracted events relevant to the apoptosis pathway to gain insight into the biological relevance of the result. The dataset, consisting of 19.2 million occurrences of 4.5 million unique events, is freely available for use in research at http://bionlp.utu.fi/.

1 Introduction

There has recently been substantial interest in *event models* in biomedical information extraction (IE). The expressive event representation captures extracted knowledge as structured, recursively nested, typed associations of arbitrarily many participants in specific roles. The BioNLP'09 Shared Task on Event Extraction (Kim et al., 2009), the first large scale evaluation of biomedical event extraction systems, drew the participation of 24 groups and established a standard event representation scheme and datasets. The training and test data of the Shared Task comprised 13,623 manually annotated events in 1,210 PubMed citation abstracts, and on this data the top performing system of Björne et al. (2009; 2010b) achieved an overall F-score of 51.95% (Kim et al., 2009).

The issue of the scalability and generalization ability of the introduced event extraction systems beyond the domain of the GENIA corpus on which the Shared Task was based has remained largely an open question. In a prior study, we have established on a 1% random sample of PubMed titles and abstracts that the event extraction system of Björne et al. is able to scale up to PubMed-wide extraction without prohibitive computational time requirements, however, the actual extraction from the entire PubMed was left as a future work (Björne et al., 2010a). Thus, the top-ranking event extraction systems in the Shared Task have, in fact, not been used so far for actual mass-scale event extraction beyond the carefully controlled setting of the Shared Task itself. Further, since an automated named entity recognition step was not part of the Shared Task, the interaction of the event extraction systems with gene/protein name recognizers remains largely unexplored as well.

In this study, we address some of these questions by performing a mass-scale event extraction experiment using the best performing system[1] of the Shared Task (Björne et al., 2009; Björne et al., 2010b), and applying it to the entire set of titles and abstracts of the nearly 18 million citations in the 2009 distribution of PubMed. The extraction result, containing 19.2 million event occurrences, is the largest dataset of its type by several orders of magnitude and arguably represents the state-of-the-art in automatic event extraction with respect to both accuracy and size.

To support emerging community efforts in tasks that build on event extraction output, such as event network refinement, hypothesis generation, pathway extraction, and others, we make the entire resulting dataset freely available for research purposes. This allows researchers interested in questions involving *text mining*, rather than initial in-

[*] Equal contribution by first three authors.

[1] Available at http://bionlp.utu.fi/

Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, ACL 2010, pages 28–36,
Uppsala, Sweden, 15 July 2010. ©2010 Association for Computational Linguistics

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

Table 1: Targeted event types with brief example statements expressing an event of each type. In the examples, the word or words marked as triggering the presence of the event are shown in italics and event participants underlined. The event types are grouped by event participants, with the first five types taking one *theme*, binding events taking multiple *theme*s and the regulation types *theme* and *cause* participants. Adapted from (Björne et al., 2009).

formation extraction, to make use of the many favorable statistical properties of the massive dataset without having to execute the laborious and time-consuming event extraction pipeline.

In the following, we describe the Shared Task event representation applied throughout this study, the event extraction pipeline itself, and a first set of analyzes of multiple aspects of the resulting dataset.

2 Event extraction

The event extraction pipeline follows the model of the BioNLP'09 Shared Task in its representation of extracted information. The primary extraction targets are gene or gene product-related entities and nine fundamental biomolecular event types involving these entities (see Table 1 for illustration).

Several aspects of the event representation, as defined in the context of the Shared Task, differentiate the event extraction task from the body of domain IE studies targeting e.g. protein–protein interactions and gene–disease relations, including previous domain shared tasks (Nédellec, 2005; Krallinger et al., 2008). Events can have an arbitrary number of participants with specified roles (e.g. *theme* or *cause*), making it possible to capture n-ary associations and statements where some participants occur in varying roles or are only occasionally mentioned. A further important property is that event participants can be other events, resulting in expressive, recursively nested structures. Finally, events are given GENIA Event ontology types drawn from the community-standard Gene Ontology (The Gene Ontology Consortium, 2000), giving each event well-defined semantics.

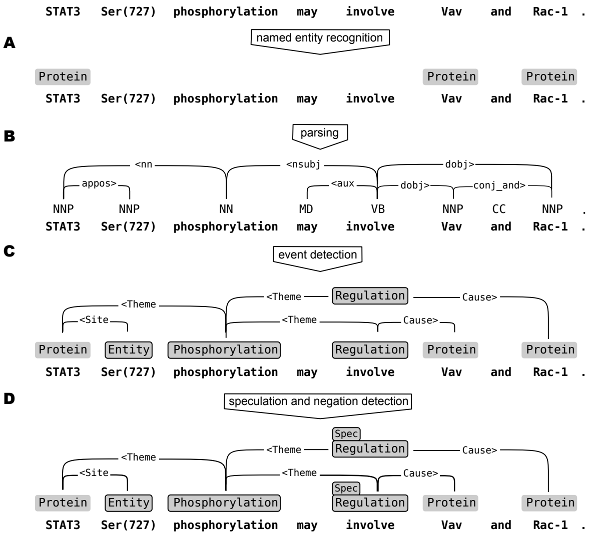
STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .
A
named entity recognition
Protein
Protein
Protein
STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .
B
parsing
<nn
<nsubj
dobj>
appos>
<aux
dobj>
conj_and>
NNP NNP NN MD VB NNP CC NNP .
STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .
C
event detection
<Theme
Regulation
Cause>
<Theme
<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>
<Theme
<Site
<Theme
Cause>
Spec
Protein Entity Phosphorylation Regulation Protein Protein
STAT3 Ser(727) phosphorylation may involve Vav and Rac-1 .

Figure 1: Event extraction. A multi-stage system produces an *event graph* for each sentence. Named entities are detected (A) using BANNER. Independently of named entity detection, sentences are parsed (B) to produce a dependency parse. Event detection (C) uses the named entities and the parse in predicting the *trigger nodes* and *argument edges* that form the events. Finally, polarity and certainty (D) are predicted for the generated events. Adapted from (Björne et al., 2009).

2.1 Event Extraction Pipeline

The event extraction pipeline applied in this work consists of three main processing steps: named entity recognition, syntactic parsing, and event extraction. The process is illustrated in Figure 1.

For named entity recognition, we use the BANNER system of Leaman and Gonzales (2008), which in its current release achieves results close to the best published on the standard GENETAG dataset and was reported to have the best performance in a recent study comparing publicly available taggers (Kabiljo et al., 2009). Titles and abstracts of all 17.8M citations in the 2009 distribution of PubMed are processed through the BANNER system.

Titles and abstracts of PubMed citations in which at least one named entity was identified, and which therefore contain a possible target for event extraction, are subsequently split into sentences using a maximum-entropy based sentence splitter trained on the GENIA corpus (Kazama and Tsujii, 2003) with limited rule-based post-processing for some common errors.

All sentences containing at least one named entity are then parsed with the domain-adapted McClosky-Charniak parser (McClosky and Charniak, 2008; McClosky, 2009), which has achieved the currently best published performance on the GENIA Treebank (Tateisi et al., 2005). The constituency parse trees are then transformed to the *collapsed-ccprocessed* variant of the Stanford Dependency scheme using the conversion tool[2] introduced by de Marneffe et al. (2006).

Finally, events are extracted using the Turku Event Extraction System of Björne et al. which achieved the best performance in the BioNLP'09 Shared Task and remains fully competitive with even the most recent advances (Miwa et al., 2010). We use a recent publicly available revision of the event extraction system that performs also extraction of Shared Task subtask 2 and 3 information, providing additional event arguments relevant to event sites and localization (*site*, *atLoc*, and *toLoc* role types in the Shared Task) as well as information on event polarity and certainty (Björne et al., 2010b).

2.2 Extraction result and computational requirements

Named entity recognition using the BANNER system required in total roughly 1,800 CPU-hours and resulted in 36,454,930 named entities identified in 5,394,350 distinct PubMed citations.

Parsing all 20,037,896 sentences with at least one named entity using the McClosky-Charniak parser and transforming the resulting constituency trees into dependency analyzes using the Stanford conversion tool required about 5,000 CPU-hours, thus averaging 0.9 sec per sentence. Even though various stability and scalability related problems were met during the parsing process, we were able to successfully parse 20,020,266 (99.91%) of all sentences.

Finally, the event extraction step required approximately 1,500 CPU-hours and resulted in 19,180,827 event instances. In total, the entire corpus of PubMed titles and abstracts was thus processed in roughly 8,300 CPU-hours, or, 346 CPU-days, the most time-consuming step by far being the syntactic parsing.

We note that, even though the components used in the pipeline are largely well-documented and mature, a number of technical issues directly related to, or at least magnified by, the untypically large dataset were met at every point of the pipeline. Executing the pipeline was thus far from a trivial undertaking. Due to the computational requirements of the pipeline, cluster computing systems were employed at every stage of the process.

2.3 Evaluation

We have previously evaluated the Turku Event Extraction System on a random 1% sample of PubMed citations, estimating a precision of 64% for event types and arguments pertaining to subtask 1 of the Shared Task (Björne et al., 2010a), which compares favorably to the 58% precision the system achieves on the Shared Task dataset itself (Björne et al., 2009).

To determine precision on subtasks 2 and 3 on PubMed citations, we manually evaluate 100 events with *site* and *location* arguments (subtask 2) and 100 each of events predicted to be *speculated* or *negated* (subtask 3).

Subtask 2 *site* and *location* arguments are mostly external to the events they pertain to and therefore were evaluated independently of their parent event. Their precision is 53% (53/100), comparable to the 58% precision established on the BioNLP'09 Shared Task development set, using the same parent-independent criterion.

To estimate the precision of the negation detection (subtask 3), we randomly select 100 events predicted to be negated. Of these, 9 were incorrect as events to such an extent that the correctness of the predicted negation could not be judged and, among the remaining 91 events, the negation was correctly predicted in 82% of the cases. Similarly, to estimate the precision of speculation detection, we randomly select 100 events predicted to be speculated, of which 20 could not be judged for correctness of speculation. Among the remaining 80, 88% were correctly predicted as speculative events. The negations were mostly signalled by explicit statements such as *is not regulated*, and speculation by statements, such as *was studied*, that defined the events as experimental questions.

[2] `http://www-nlp.stanford.edu/downloads/lex-parser.shtml`

For comparison, on the BioNLP'09 Shared Task development set, for correctly predicted events, precision for negation examples was 83% (with recall of 53%) and for speculation examples 77% (with recall of 51%).

In the rest of this paper, we turn our attention to the extraction result.

3 Term-NE mapping

As the event types are drawn from the Gene Ontology and the original data on which the system is trained has been annotated with reference to the GO definitions, the events targeted by the extraction system have well-defined biological interpretations. The meaning of complete event structures depends also on the participating entities, which are in the primary event extraction task constrained to be of gene/gene product (GGP) types, as annotated in the GENIA GGP corpus (Ohta et al., 2009a). The simple and uniform nature of these entities makes the interpretation of complete events straightforward.

However, the semantics of the entities automatically tagged in this work are somewhat more openly defined. The BANNER system was trained on the GENETAG corpus, annotated for "gene/protein entities" without differentiating between different entity types and marking entities under a broad definition that not only includes genes and gene products but also related entities such as gene promoters and protein complexes, only requiring that the tagged entities be specific (Tanabe et al., 2005). The annotation criteria of the entities used to train the BANNER system as well as the event extraction system also differ in the extent of the marked spans, with GENIA GGP marking the minimal name and GENETAG allowing also the inclusion of head nouns when a name occurs in modifier position. Thus, for example, the latter may annotate the spans *p53 gene*, *p53 protein*, *p53 promoter* and *p53 mutations* in contexts where the former would in each case mark only the substring *p53*.

One promising future direction for the present effort is to refine the automatically extracted data into an event network connected to specific entries in gene/protein databases such as Entrez Gene and UniProt. To achieve this goal, the resolution of the tagged entities can be seen to involve two related but separate challenges. First, identifying the specific database entries that are referred to by the genes/proteins named in the tagged entities, and second, mapping from the events involving automatically extracted terms to ones involving the associated genes/proteins. The first challenge, gene/protein name normalization, is a well-studied task in biomedical NLP for which a number of systems with promising performance have been proposed (Morgan and Hirschman, 2007). The second we believe to be novel. In the following, we propose a method for resolving this task.

Relation	Examples
Equivalent	GGP gene, wild-type GGP
Class-Subclass	human GGP, HIV-1 GGP
Object-Variant	
GGP-Isoform	GGP isoform
GGP-Mutant	dominant-negative GGP
GGP-Recombinant	GGP expression plasmid
GGP-Precursor	GGP precursor, pro-GGP
Component-Object	
GGP-Amino acid	GGP-Ile 729
GGP-AA motif	GGP NH2-terminal
GGP-Reg. element	GGP proximal promoter
GGP-Flanking region	GGP 5' upstream sequence
Object-Component	
GGP-Protein Complex	GGP homodimers
Place-Area	
GGP-Locus	GGP loci
Member-Collection	
GGP-Group	GGP family members

Table 2: Gene/gene product NE-term relation types with examples. Top-level relations in the relation type hierarchy shown in bold, specific NE names in examples replaced with *GGP*. Intermediate levels in the hierarchy and a number of minor relations omitted. Relation types judged to allow remapping (see text) underlined.

We base the decision on how to map events referencing broadly defined terms to ones referencing associated gene/protein names in part on a recently introduced dataset of "static relations" (Pyysalo et al., 2009) between named entities and terms (Ohta et al., 2009b). This dataset was created based on approximately 10,000 cases where GGP NEs, as annotated in the GENIA GGP corpus (Ohta et al., 2009a), were embedded in terms, as annotated in the GENIA term corpus (Ohta et al., 2002). For each such case, the relation between the NE and the term was annotated using a set of introduced relation types whose granularity was defined with reference to MeSH terms (see Table 2, Ohta et al., 2009b). From this data, we extracted prefix and suffix strings that, when affixed to a GGP name, produced a term with a predictable relation (within the dataset) to the GGP. Thus, for example, the

term	GGP
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
p24 antigen	p24
tr. factor NF-kappaB	NF-kappaB
MHC molecules	MHC
PKC isoforms	PKC
HLA alleles	HLA
RET proto-oncogene	RET
ras oncogene	ras
SV40 DNA	SV40
EGFR tyrosine kinase	EGFR

Table 3: Examples of frequently applied mappings. Most frequent term for each mapping is shown. Some mention strings are abbreviated for space.

	Mentions	Types
Total	36454930	4747770
Mapped	2212357 (6.07%)	547920 (11.54%)
Prefix	430737 (1.18%)	129536 (2.73%)
Suffix	1838646 (5.04%)	445531 (9.38%)

Table 4: Statistics for applied term-GGP mappings. Tagged mentions and types (unique mentions) shown separately. Overall total given for reference, for mappings overall for any mapping shown and further broken down into prefix-string and suffix-string based.

prefix string “wild-type” was associated with the *Equivalent* relation type and the suffix string “activation sequence” with the *GGP-Regulatory element* type. After filtering out candidates shorter than 3 characters as unreliable (based on preliminary experiments), this procedure produced a set of 68 prefix and 291 suffix strings.

To make use of the data for predicting relations between GGP names and the terms formed by affixing a prefix or suffix string, it is necessary to first identify name-term pairs. Candidates can be generated simply by determining the prefix/suffix strings occurring in each automatically tagged entity and assuming that what remains after removing the prefixes and suffixes is a GGP name. However, this naive strategy often fails: while removing “protein” from “p53 protein” correctly identifies “p53” as the equivalent GGP name, for “capsid protein” the result, “capsid” refers not to a GGP but to the shell of a virus – “protein” is properly part of the protein name. To resolve this issue, we drew on the statistics of the automatically tagged entities, assuming that if a prefix/suffix string is not a fixed part of a name, the name will appear tagged also without that string. As the tagging covers the entire PubMed, this is likely to hold for all but the very rarest GGP names. To compensate for spurious hits introduced by tagging errors, we specifically required that to accept a candidate prefix/suffix string-name pair, the candidate name should occur more frequently without the prefix/suffix than with it. As the dataset is very large, this simple heuristic often gives the right decision with secure margins: for example, “p53” was tagged 117,835 times but “p53 protein” only 11,677, while “capsid” was (erroneously) tagged 7 times and “capsid protein” tagged 1939 times.

A final element of the method is the definition of a mapping to events referencing GGP NEs from the given events referencing terms, the NEs contained in the terms, and the NE-term relations. In this work, we apply independently for each term a simple mapping based only on the relation types, deciding for each type whether replacing reference to a term with reference to a GGP holding the given relation to the term preserves event semantics (to an acceptable approximation) or not. For the Equivalent relation this holds by definition. We additionally judged all Class-Subclass and Component-Object relations to allow remapping (accepting e.g. P_1 *binds part of* P_2 → P_1 *binds* P_2) as well as selected Object-Variant relations (see Table 2). For cases judged not to allow remapping, we simply left the event unmodified.

Examples of frequently applied term-GGP mappings are shown in Table 3, and Table 4 shows the statistics of the applied mappings. We find that suffix-based mappings apply much more frequently than prefix-based, perhaps reflecting also the properties of the source dataset. Overall, the number of unique tagged types is reduced by over 10% by this procedure. It should be noted that the applicability of the method could likely be considerably extended by further annotation of NE-term relations in the dataset of Ohta et al. (2009b): the current data is all drawn from the GENIA corpus, drawn from the subdomain of transcription factors in human blood cells, and its coverage of PubMed is thus far from exhaustive.

4 Event recurrence

Given a dataset of events extracted from the entire PubMed, we can study whether, and to what extent, events are re-stated in multiple PubMed citations. This analysis may shed some light — naturally within the constraints of an automatically extracted dataset rather than gold-standard annotation — on the often (informally) discussed hypothesis that a high-precision, low recall system might be a preferred choice for large-scale extraction as the lower recall would be compensated by the redundancy of event statements in PubMed.

In order to establish event recurrence statistics, that is, the number of times a given event is repeated in the corpus, we perform a limited normalization of tagged entities consisting of the Term-NE mapping presented in Section 3 followed by lowercasing and removal of non-alphanumeric characters. Two named entities are then considered equal if their normalized string representations are equal. For instance, the two names *IL-2 gene* and *IL2* would share the same normalized form *il2* and would thus be considered equal.

For the purpose of recurrence statistics, two events are considered equal if their types are equal, and all their Theme and Cause arguments, which can be other events, are recursively equal as well. A canonical order of arguments is used in the comparison, thus e.g. the following events are considered equal:

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

In total, the system extracted 19,180,827 instances of 4,501,883 unique events. On average, an event is thus stated 4.2 times. The distribution is, however, far from uniform and exhibits the "long tail" typical of natural language phenomena, with 3,484,550 (77%) of events being singleton occurrences. On the other hand, the most frequent event, *localization(Theme:insulin)*, occurs as many as 59,821 times. The histogram of the number of unique events with respect to their occurrence count is shown in Figure 2.

The total event count consists mostly of simple one-argument events. The arguably more interesting category of events that involve at least two different named entities constitutes 2,064,278 instances (11% of the 19.2M total) of 1,565,881 unique events (35% of the 4.5M total). Among these complex events, recurrence is thus considerably lower, an event being stated on average 1.3 times. The most frequent complex event, with 699 occurrences, is *positive-regulation(Cause:GnRG,Theme:localization(Theme:LH))*, reflecting the well-known fact that *GnRG* causes the release of *LH*, a hormone important in human reproduction.

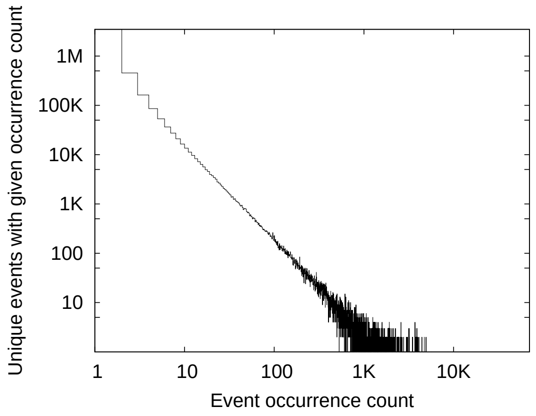

Figure 2: Number of unique events (y-axis) with a given occurrence count (x-axis).

	R	P	N	L	B	E	T	C	H
R	**561**	173	128	42	63	83	30	16	17
P	173	**1227**	192	58	99	143	39	20	23
N	128	192	**668**	46	73	98	31	17	18
L	42	58	46	**147**	57	75	25	15	15
B	63	99	73	57	**1023**	134	35	20	21
E	83	143	98	75	134	**705**	49	22	24
T	30	39	31	25	35	49	**79**	11	11
C	16	20	17	15	20	22	11	**39**	7
H	17	23	18	15	21	24	11	7	**49**

Table 5: Event type confusion matrix. Each element contains the number of unique events, in thousands, that are equal except for their type. The matrix is symmetric and its diagonal sums to 4,5M, the total number of extracted unique events. The event types are (R)egulation, (P)ositive regulation, (N)egative regulation, (L)ocalization, (B)inding, gene (E)xpression, (T)ranscription, protein (C)atabolism, and p(H)osphorylation.

To gain an additional broad overview of the characteristics of the extracted events, we compute an *event type confusion matrix*, shown in Table 5. In this matrix, we record for each pair of event types T_1 and T_2 the number of unique events of type T_1 for which an event of type T_2 can be found such that, apart for the type difference, the events are otherwise equal. While e.g. a positive regulation-negative regulation pair is at least unusual, in general these event pairs do not suggest extraction errors: for instance the existence of the event *expression(Theme:A)* does not in any way prevent the existence of the event *localization(Theme:A)*, and *regulation* subsumes *positive-regulation*. Nevertheless, Table 5 shows a clear preference for a single type for the events.

5 Case Study: The apoptosis pathway

In this section, we will complement the preceding broad statistical overview of the extracted events with a detailed study of a specific pathway, the *apoptosis pathway*, determining how well the extracted events cover its interactions (Figure 3).

To create an event network, the events must be linked through their protein arguments. In addition to the limited named entity normalization introduced in Section 4, we make use of a list of synonyms for each protein name in the apoptosis pathway, obtained manually from protein databases, such as UniProt. Events whose protein arguments correspond to any of these known synonyms are then used for reconstructing the pathway.

The apoptosis pathway consists of several overlapping signaling routes and can be defined on different levels of detail. To have a single, accurate and reasonably high-level definition, we based our pathway on a concisely presentable subset of the KEGG human apoptosis pathway (entry hsa04210) (Kanehisa and Goto, 2000). As seen in Figure 3, the extracted dataset contains events between most interaction partners in the pathway.

The constructed pathway also shows that the extracted events are not necessarily interactions in the physical sense. Many “higher level” events are extracted as well. For example, the extracellular signaling molecule *TNF*α can trigger pathways leading to the activation of *Nf-*κ*B*. Although the two proteins are not likely to interact directly, it can be said that *TNF*α upregulates *NF-*κ*B*, an event actually extracted by the system. Such statements of indirect interaction co-exist with statements of actual, physical interactions in the event data.

6 Conclusions

In this paper, we have presented the result of processing the entire, unabridged set of PubMed titles and abstracts with a state-of-the-art event extraction pipeline as a new resource for text mining in the biomedical domain. The extraction result arguably represents the best event extraction output achievable with currently available tools.

The primary contribution of this work is the set of over 19M extracted event instances of 4.5M unique events. Of these, 2.1M instances of 1.6M unique events involve at least two different named entities. These form an event network several orders of magnitude larger than those previously available. The data is intended to support research in biological hypothesis generation, pathway extraction, and similar higher-level text mining tasks. With the network readily available in an easy-to-process format under an open license, researchers can focus on the core tasks of text mining without the need to perform the tedious and computationally very intensive task of event extraction with a complex IE pipeline.

In addition to the extracted events, we make readily available the output of the BANNER system on the entire set of PubMed titles and abstracts as well as the parser output of the McClosky-Charniak domain-adapted parser (McClosky and Charniak, 2008; McClosky, 2009) further transformed to the Stanford Dependency representation using the tools of de Marneffe et al. (2006) for nearly all (99.91%) sentences with at least one named entity identified. We expect this data to be of use for the development and application of systems for event extraction and other BioNLP tasks, many of which currently make extensive use of dependency syntactic analysis. The generation of this data having been far from a trivial technical undertaking, its availability as-is can be expected to save substantial duplication of efforts in further research.

A manual analysis of extracted events relevant to the apoptosis pathway demonstrates that the event data can be used to construct detailed biological interaction networks with reasonable accuracy. However, accurate entity normalization, in particular taking into account synonymous names, seems to be a necessary prerequisite and remains among the most important future work directions. In the current study, we take first steps in this direction in the form of a term-NE mapping method in event context. The next step will be the application of a state-of-the-art named entity normalization system to obtain biological database identities for a number of the named entities in the extracted event network, opening possibilities for combining the data in the network with other biological information. A further practical problem to address will be that of visualizing the network and

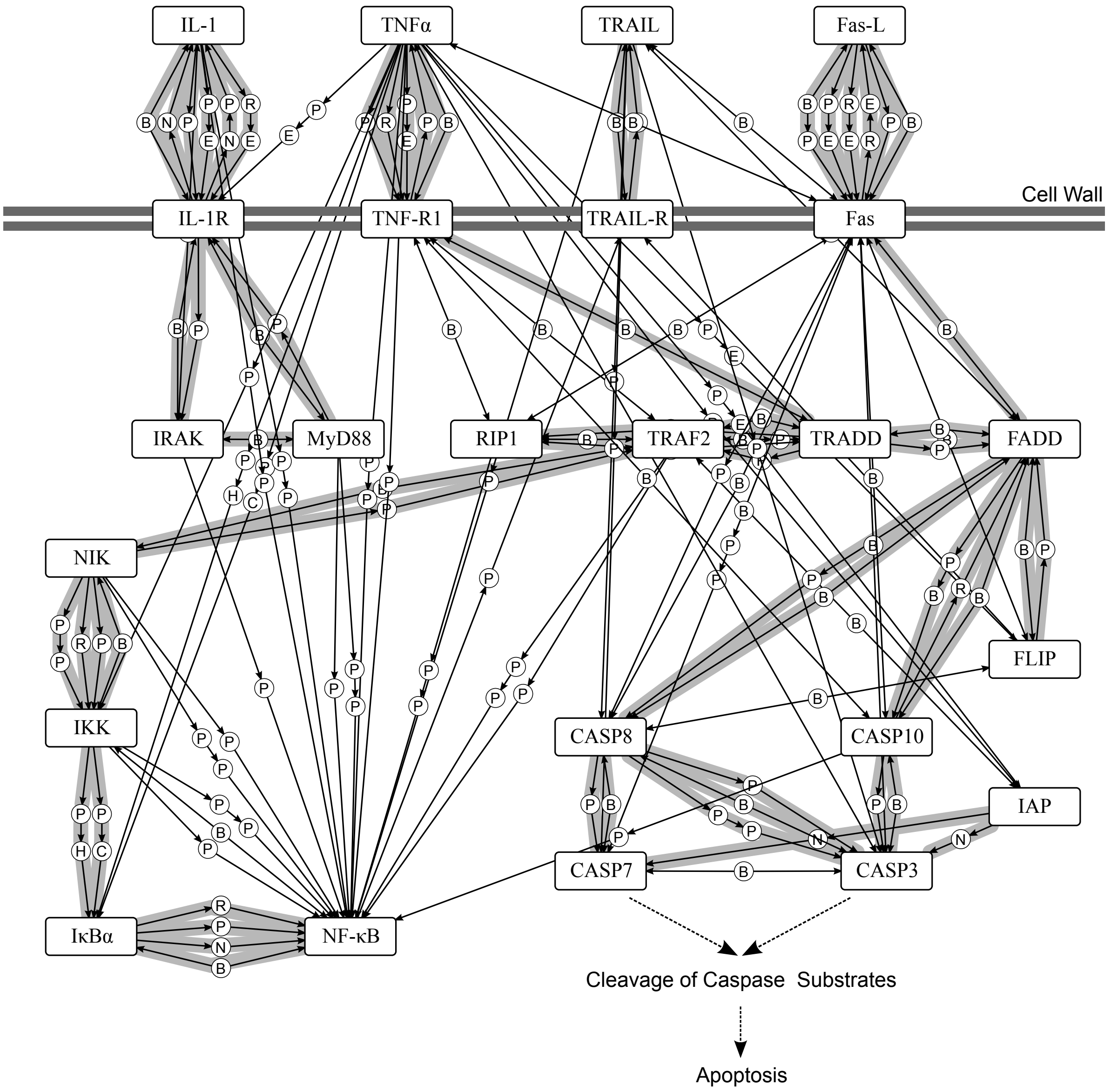

Figure 3: Extracted apoptosis event network. Events shown in the figure are selected on their prominence in the data or correspondence to known apoptosis interactions. Events corresponding to KEGG apoptosis pathway interaction partners are highlighted with a light grey background. The event types are (P)ositive regulation, (N)egative regulation, (R)egulation, gene (E)xpression, (B)inding, p(H)osphorylation, (L)ocalization and protein (C)atabolism.

presenting the information in a biologically meaningful manner.

The introduced dataset is freely available for research purposes at `http://bionlp.utu.fi/`.

Acknowledgments

This work was supported by the Academy of Finland and by Grant-in-Aid for Specially Promoted Research (MEXT, Japan). Computational resources were provided by *CSC – IT Center for Science, Ltd.*, a joint computing center for Finnish academia and industry. We thank Robert Leaman for advance access and assistance with the newest release of BANNER.

References

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010a. Complex event extraction at PubMed scale. In *Proceedings of the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2010)*. In press.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2010b. Extracting contextualized complex biological events with rich graph-based feature sets. *Computational Intelligence*. In press.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC-06*, pages 449–454.

Renata Kabiljo, Andrew Clegg, and Adrian Shepherd. 2009. A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics*, 10(1):233.

M. Kanehisa and S. Goto. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, Jan.

Jun'ichi Kazama and Jun'ichi Tsujii. 2003. Evaluation and extension of maximum entropy models with inequality constraints. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 137–144.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 1–9, Boulder, Colorado. ACL.

Martin Krallinger, Florian Leitner, Carlos Rodriguez-Penagos, and Alfonso Valencia. 2008. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4.

R. Leaman and G. Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing*, pages 652–663.

David McClosky and Eugene Charniak. 2008. Self-Training for Biomedical Parsing. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT'08)*, pages 101–104.

David McClosky. 2009. *Any Domain Parsing: Automatic Domain Adaptation for Natural Language Parsing*. Ph.D. thesis, Department of Computer Science, Brown University.

Makoto Miwa, Rune Sætre, Jin-Dong Kim, and Jun'ichi Tsujii. 2010. Event Extraction With Complex Event Classification Using Rich Features. *J Bioinform Comput Biol*, 8:131–146.

Alexander A. Morgan and Lynette Hirschman. 2007. Overview of BioCreative II gene normalization. In *Proceedings of BioCreative II*, pages 101–103.

Claire Nédellec. 2005. Learning Language in Logic - Genic Interaction Extraction Challenge. In J. Cussens and C. Nédellec, editors, *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, pages 31–37.

Tomoko Ohta, Yuka Tateisi, Hideki Mima, and Jun'ichi Tsujii. 2002. GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the Human Language Technology Conference (HLT'02)*, pages 73–77.

Tomoko Ohta, Jin-Dong Kim, Sampo Pyysalo, Yue Wang, and Jun'ichi Tsujii. 2009a. Incorporating genetag-style annotation to genia corpus. In *Proceedings of the BioNLP 2009 Workshop*, pages 106–107, Boulder, Colorado, June. Association for Computational Linguistics.

Tomoko Ohta, Sampo Pyysalo, Kim Jin-Dong, and Jun'ichi Tsujii. 2009b. A re-evaluation of biomedical named entity - term relations. In *Proceedings of LBM'09*.

Sampo Pyysalo, Tomoko Ohta, Jin-Dong Kim, and Jun'ichi Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the BioNLP 2009 Workshop*, pages 1–9, Boulder, Colorado, June. Association for Computational Linguistics.

Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. 2005. GENETAG: A tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6(Suppl. 1):S3.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax Annotation for the GENIA corpus. In *Proceedings of the IJCNLP 2005, Companion volume*, pages 222–227.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25:25–29.