Event extraction across multiple levels of biological organization

Sampo Pyysalo\textsuperscript{1}, Tomoko Ohta\textsuperscript{1}, Makoto Miwa\textsuperscript{1}, Han-Cheol Cho\textsuperscript{2}, Jun'ichi Tsujii\textsuperscript{3} and Sophia Ananiadou\textsuperscript{1}

\textsuperscript{1}National Centre for Text Mining and School of Computer Science, University of Manchester, Manchester, UK, \textsuperscript{2}Department of Computer Science, University of Tokyo, Tokyo, Japan and \textsuperscript{3}Microsoft Research Asia, Beijing, China

ABSTRACT

Motivation: Event extraction using expressive structured representations has been a significant focus of recent efforts in biomedical information extraction. However, event extraction resources and methods have so far focused almost exclusively on molecular-level entities and processes, limiting their applicability.

Results: We extend the event extraction approach to biomedical information extraction to encompass all levels of biological organization from the molecular to the whole organism. We present the ontological foundations, target types and guidelines for entity and event annotation and introduce the new multi-level event extraction (MLEE) corpus, manually annotated using a structured representation for event extraction. We further adapt and evaluate named entity and event extraction methods for the new task, demonstrating that both can be achieved with performance broadly comparable with that for established molecular entity and event extraction tasks.

Availability: The resources and methods introduced in this study are available from http://nactem.ac.uk/MLEE/.

Contact: Pyysalo@cs.man.ac.uk

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

A detailed understanding of biological systems requires the ability to trace cause and effect across multiple levels of biological organization, from molecular-level reactions to cellular, tissue- and organ-level effects to organism-level outcomes (Kerr et al. 2002). Consequently, any effort aiming to comprehensively represent biological systems must address entities and processes at all of these levels.

This challenge has so far been only partially met in biomedical information extraction (IE) and text mining, which aim to improve access to domain knowledge by automating aspects of processing the literature. Until recently, efforts in domain IE were primarily focused on the basic task of recognizing mentions of relevant entities such as genes and proteins in text (Matsubara et al. 2008) and on the extraction of pairwise relations between these representatives, for example, protein–protein interactions (Kull et al. 2007, Nodelski 2009). Such representations lack the capacity to capture any but the simplest of associations.

In recent years, there has been increasing interest in the extraction of structured representations capable of capturing associations of arbitrary numbers of participants in specific roles. Such approaches to IE, frequently termed event extraction, are capable of representing complex associations—such as the binding of a protein to another inhibiting its localization to a specific cellular compartment (Fig. 1)—and open many new opportunities for domain text mining applications ranging from semantic search to database and pathway curation support (Ananiadou et al. 2010). There is significant momentum behind the move to richer representations for IE: more than 30 groups have introduced methods for biomedical event extraction in shared tasks (Kim et al. 2011a, b); event-annotated corpora have been introduced for many extraction targets, including DNA methylation (Ohta et al. 2011a), protein modifications (Pyysalo et al. 2011) and the molecular mechanisms of infectious diseases (Pyysalo et al. 2012); event extraction methods have been applied to automatically analyze all 20 million PubMed abstracts (Björne et al. 2011a); and event extraction analyses are being integrated into literature search systems such as MEDIE and applied in support of advanced tasks such as pathway curation (Ohta et al. 2011c).

While the event extraction approach has been demonstrated to be applicable to a variety of extraction targets across different subdomains of biomedical science, related efforts all share a key restriction: nearly exclusive focus on molecular-level entities and events. Entities such as proteins and genes and events such as binding and phosphorylation are an important part of the picture of biological systems, but still only a part, and any IE approach aiming to capture the whole picture must also consider other levels of biological organization.

In this study, our aim is to extend the scope of existing event extraction resources and methods to levels of biological organization ranging from the subcellular to the organism level as a step toward developing the capacity for the automatic extraction of these targets from the entire available literature. Toward this end, we propose relevant entity and event types for annotation across these levels with reference to community-standard ontologies, develop a set of detailed guidelines for their annotation in text and create structured event annotation marking over 8000 entities and 6000 events in abstracts relevant to cancer biology, previously annotated by domain experts to identify spans of text relevant to their interests. Using this data, we perform experiments using state-of-the-art methods for both entity mention detection and event extraction to analyze

\textsuperscript{1}http://www.nactem.ac.uk/mde/.

\textsuperscript{2}Some recent tasks have considered also organisms (primarily unicellular, see e.g. Bossy et al. 2005, Pyysalo et al. 2011).
We apply the specific event representation first formalized in

the feasibility of extraction using existing tools, further evaluating

the benefits of specific adaptations of such tools to the novel task.

2 APPROACH

2.1 Corpus texts and reference annotation

We selected as the starting point for our study a recently introduced
corpus of 262 PubMed abstracts on angiogenesis, the development
of new blood vessels from existing ones. The domain involves

Figure 1: Example sentence with event annotation. PROT. - REG and CELL
comp. abbreviated for PROTEIN, NEGATIVE REGULATION and CELL COMPONENT,
respectively.

2.2 Representation

We apply the specific event representation first formalized in
the BioNLP 2009 Shared Task on event extraction and applied
in numerous resources and methods introduced since. In this
representation, Entity mentions (or entities, for short) are marked as
continuous spans of text identified with a type (e.g. PROTEIN), and event
structures (or events) are n-ary associations of participants—
entities or other events—each of which is identified as participating
in the event in a specific role (e.g. Theme and Cause). Each event is
assigned a type from a fixed set defined for the task (e.g. BINDING
and PHOSPHORYLATION) and is associated with a specific span of text
stating the event, termed the event trigger. Events can additionally be
marked with modifiers identifying the event as being, e.g. explicitly
negated, or stated in a speculative context. We refer to Kim et al
(2011) for a detailed presentation of the representation.

Figure 2: Span versus structure. Although a representation using nested, typed
spans (left) can capture the fact that specific entities participate in a process,
it lacks the mechanisms to express, e.g. the direction of causality. The
structured event representation (right) differentiates Themes from Causes

2.3 Ontological basis

We take as basic the division between continuants (or endurants)
and occurrents (perdurants, processes or events) (see e.g.
Smith, 2003) and adopt the general principle followed also in major
previously introduced event-annotated resources that references
to continuants such as material entities are annotated using the
entity representation and references to occurrents such as biological
processes are annotated as events.

In the definition of our annotation scheme, we aim for
compatibility with existing event-annotated corpora—primarily the
five ‘main task’ corpora introduced in the BioNLP Shared Tasks—
to allow these to be used together with the annotations that we
create and to assure that our extensions are coherent with existing
resources derived from these corpora. Thus, for molecular-level
entity and process types, we adopt the scope, semantics and
annotation guidelines of these resources as closely as possible
without compromising coverage of mentions marked as relevant
by domain experts. For entities and processes not in scope of
previous event resources, we propose new types for annotation,
basing type and scope definitions and annotation guidelines on major
community-curated ontological resources from the open biomedical
ontologies (OBO) foundry4 Smith et al. (2009). In brief, before
primary annotation, we analyzed mentions marked in the reference
annotation to identify entity and process types not in scope of
previously defined event annotation guidelines and then defined
new types and guidelines for annotation with reference to selected
ontologies. These are summarized in the following.

2.4 Annotation scheme

The focus on our extensions of previously proposed event annotation
schemes is on anatomical entities such as cells, tissues and organs
and processes involving them such as growth, remodeling and death.

For anatomical entity types, we adopt a top-level division by
granularity Krämer et al. (2008) based primarily on the upper-
level structure of the Common Anatomy Reference Ontology
(CARO) Haendel et al. (2008), an organism-independent ontology
of anatomy based on the human-specific Foundational Model of
Anatomy Rouse and Meijers (2003, 2008), as outlined in our
previous work on anatomical entities Pyysalo et al. (2011b). To
account for pathological anatomy-level entities (e.g. glioma)—out of

4 We use the terms ‘entity’ and ‘event’ primarily following usage in IE, to
identify forms of representation, not ontological categories. In particular, the
latter term does not denote a category distinct from processes.
http://obofoundry.org.

5 Although the existing corpus annotation of Wang et al. (2010) identifies
such mentions, they are typed nonspecifically, using e.g. POSITIVE
regulation to mark ‘development’ and NEGATIVE regulation for ‘cell’
death.

Copied by: SK
MANUSCRIPT CATEGORY: ECCB

Page: 1576

Downloaded from https://academic.oup.com/bioinformatics/article-abstract/28/18/i575/249872
By guest
on 28 July 2018

[12:52 6/8/2012 Bioinformatics-bts407.tex]

Page: 1576

i576
Table 1. Primary entity types, related ontology terms and annotation counts

| Type                      | Term(s)                                      | Examples                                      | Count |
|---------------------------|----------------------------------------------|-----------------------------------------------|-------|
| ORGANISM                  | Single cell org.,multi-cellular org.         | Human, mouse, C. albidus                      | 722   |
| ORGANISM subdivision      | Organism subdivision                        | Head, thorax, hindlimb, legs                 | 49    |
| ANATOMICAL SYSTEM         | Anatomical system                           | Central nervous system, pulmonary system     | 18    |
| ORGAN                     | Compound organ                              | Heart, eyes, skin                            | 176   |
| MULTITISSUE STRUCTURE     | Multi-tissue structure                       | Blood vessel, perivascular membrane, lymph nodes | 554  |
| TISSUE                    | Portion of tissue                           | Endothelium, adipose tissue, capillary       | 426   |
| CELL                      | Cell                                         | Endothelial cells, HUVECs, pericyte, cancer cells | 1198  |
| CELLULAR COMPONENT        | Cellular component                          | Nuclei, focal adhesions, extracellular matrix | 145   |
| DEVELOPING ANATOMIC STRUCTURE | Developing anatomical structure             | Embryos                                      | 6     |
| ORGANISM SUBSTANCE        | Portion of organ substance                  | Blood, serum, plasma, urine                  | 142   |
| IMMATERIAL ANOMALOUS ENTITY | Immaterial anatomical entity            | Lumen, preprostational space, marrow cavity | 15    |
| PATHOLOGICAL FORMATION    | Cancer; benign neoplasmic                    | Tumor, colorectal cancer, gliomas            | 910   |
| MOLECULE                  | Inorganic molecular entity; drug             | Oxygen, ethanol, broca/chum, thalidomide     | 944   |
| GENE OR GENE PRODUCT      | Gene, RNAs, proteins                       | VEGF P.III, 1K, endostatin, thrombin         | 2962  |

Labels in grey identify informal categories used in evaluation.

6 Annotated also in previously introduced event extraction resources. t, identifies a term t in an ontology o; ontology identifiers are OBO Foundry prefixes (namespaces).

Scope of ontologies of canonical anatomy—we draw on the approach proposed by (Smith et al., 2005). Table 1 summarizes the primary entity types applied in the annotation.

For event types, we draw primarily on the biological process nomenclature of the gene ontology (GO) (Ashburner et al., 2000). As in previous event-annotated resources, we consider only general upper-level GO terms such as growth, to specific processes included in GO through composite terms such as regulation of heart growth, are captured using the explicitly structured representation of event concepts (Fig. 3). We also capture general statements of causal association using Regulation types, as in previous event annotation efforts (see e.g. Kim et al., 2008). Following the scope of the reference annotation, we introduce event annotation also for intentionally planned processes (e.g. injection) as outlined in the Ontology for Biomedical Investigations (OBI) (Brinkman et al., 2008), using a single, non-specific type Planned process for their annotation. We additionally introduce a Breakdown event for annotating pathological processes that result in the breakdown of anatomical structures. Finally, we apply the domain-specific Blood vessel development type to annotate references to blood vessel development through expressions such as ‘angiogenesis’ that incorporate both the process and the affected entity. Expressions such as ‘blood vessel development’ that allow explicitly structured annotation are marked with a separate entity annotation (e.g. ‘blood vessel’) and an event (e.g. ‘development’) taking the entity as its Theme. The primary event types are summarized in Table 1.

For event participants, we apply otherwise standard roles included also in previous events (e.g. Theme and Cause) but introduce the role Instrument for distinguishing entities used to carry out planned processes from those that undergo the effects of the process. Also as in previously introduced event corpora, we apply two binary modifiers, Negation and Speculation, marking events as explicitly negated (e.g. ‘cells did not proliferate’) or stated in a speculative context (e.g. ‘growth might be inhibited’), respectively.

We refer to the detailed annotation guidelines (Prevalo et al., 2013) for specifics of the annotation, but note here one systematic difference between our annotation and the scope of the reference ontologies: the ontologies define idealized types of canonical anatomy and physiological processes—but texts primarily refer to real-world instances that do not fill these exacting criteria (Iida and Hunter, 2011). We thus interpreted the scope of mentions marked with a specific type to include not only the corresponding (canonical) types defined in ontologies but also variants such as entities or processes influenced by mutation, including also pathological variants. As specific examples, we mark ‘cancer cell’ as Cells., and ‘[cancer] growth’ as Growth.

2.5 Annotation process

Primary annotation was performed by a PhD biologist with more than a decade of experience in text annotation who had previously coordinated several event annotation efforts (TO). Annotations were made using the IRAT rapid annotation tool (Stenetorp et al., 2011).

Footnotes:

6 Note that we differentiate between types applied in annotation and their (broadly) corresponding ontology types.

7 This annotation strategy can be viewed as partly analogous to efforts to make GO term structure explicit (Muscasell et al., 2010).

8 For example in ‘rats were injected with hyperforin’, the ORGANISM MENTION (‘rats’) is the Theme of the PLANNED PROCESS (‘injected’) and the DRUG OR COMPOUND MENTION (‘hyperforin’) is the Instrument.
Table 2. Primary event types, argument roles, related ontology terms and annotation counts

| Type | Arguments | Terms(s) | Examples | Count |
|------|-----------|----------|----------|-------|
| **ANATOMICAL** | | | | |
| Cell proliferation | Theme | Cell proliferation<sub>em</sub> | proliferating [ECs], [MCSs] accumulated | 133 |
| Development | Theme | Developmental process<sub>em</sub> | [skin] development, [stress fiber] formation | 316 |
| Blood vessel development | Theme, At-Loc | Blood vessel development<sub>em</sub> | angiogenesis, neovascularization | 855 |
| Growth | Theme | Growth<sub>em</sub> | growth [of arteries], [tumour] growth | 169 |
| Death | Theme | Death<sub>em</sub> | connective tissue necrosis, [cell] apoptosis | 97 |
| Breakdown | Theme | — | ECM degradation, damage [to tumour cell] (vacular) remodeling, changes [in membrane] | 33 |
| Reshaping | Theme | Tissue remodeling<sub>em</sub> | | |
| **MOLECULAR** | | | | |
| Synthesis | Theme | Biosynthetic process<sub>em</sub> | [ATP] synthesis, production [of NOS] | 17 |
| Gene expression<sup>a</sup> | Theme | Gene expression<sub>em</sub> | expression [of VEGF] | 435 |
| Transcription | Theme | Transcription, DNA-dependent<sub>em</sub> | VEGF mRNA expression | 37 |
| Catalytic<sup>b</sup> | Theme | Catabolic process<sub>em</sub> | [p53] breakdown | 26 |
| Phosphorylation<sup>c</sup> | Theme, Site | Phosphorylation<sub>em</sub> | phosphorylation [of KDR] | 33 |
| Dephosphorylation<sup>d</sup> | Theme, Site | Dephosphorylation<sub>em</sub> | [Mk-1] dephosphorylation | 6 |
| **LOCALIZATION** | Theme, At/From-To-Loc | Localization<sub>em</sub> | [VEGF] colocalized, [VEGF] wax secreted | 450 |
| Binding<sup>e</sup> | Theme, Site | Binding<sub>em</sub>, biological adhesion<sub>em</sub> | [cell] adhesion, GDP-bound [Rab5a] | 184 |
| Regulation<sup>f</sup> | Theme, Cause, Site | Biological regulation<sub>em</sub> | [AMHR] modulates [activation of AP-1] | 773 |
| Positive regulation<sup>g</sup> | Theme, Cause, Site | Pos.regulation of biol.proc<sub>em</sub> | [implant] stimulates VEGF expression | 1327 |
| Negative regulation<sup>h</sup> | Theme, Cause, Site | Neg.regulation of biol.proc<sub>em</sub> | inhibition [of NO synthase by L-NAME] | 921 |
| **PLANNED** | Theme, Instrument | Planned process<sub>em</sub> | injection [of U-995], [UFT] administration | 643 |

<i>Labels in gray identify categories used in evaluation: events of the Anatomical category involve Organism or Anatomy entities (Table 1); Molecular involve Molecule entities, others can involve any entity type.</i>

<sup>a</sup>Annotated also in previously introduced event extraction resources.

Detailed annotation guidelines were prepared based on those for the GENIA and BioNLP Shared Task guidelines and refined throughout annotation to clarify ambiguous cases and document specific decisions made in annotation. We refer to the supplementary documentation and these guidelines (Pyysalo et al., 2012) for further details of the annotation scheme and the detailed definitions of all annotated types.

3 METHODS

This section presents the automatic entity mention detection and event extraction methods applied in this study, their adaptation to the novel extraction targets and the experimental setup.

Following standard practice in domain event extraction studies, we divide the automatic extraction task into two separate stages, the detection of entity mentions and the extraction of events involving these and evaluate system performance on these two separately.

3.1 Entity mention detection

For entity mention detection experiments, we applied NERsuite, a named entity recognition toolkit based on the CRF-suite implementation by [Okazaki et al., 2007](Okazaki et al., 2007). NERsuite is capable of efficiently incorporating features based on token matching against large-scale lexical resources, and the applied version achieves an $F$ score of 86.4% on the BioCreative II evaluation standard (GENETAG) by [Tangbe et al., 2009](Tangbe et al., 2009), effectively matching the performance of the best available systems for the task.<sup>3</sup>

Following initial sentence splitting and tokenization, we perform lemmatization, POS-tagging and shallow parsing using the GENIA tagger by [Tsuruoka and Tsujii, 2005](Tsuruoka and Tsujii, 2005). Next, we optionally perform a matching step using dictionaries compiled from the UMLS Metathesaurus and OBO resource definitions. We then extract a comprehensive set of features for machine learning, building on orthographic, lexical, syntactic and dictionary match information (see Supplementary information).

Following preliminary development test experiments, we chose to apply a single model that jointly predicts all entity types. In the final experiments, we compare a base model using only from the newly annotated data without external resources with a dictionary-supported model that incorporates features from matching against the lexical resources derived from UMLS, Entrez Gene and OBO foundry ontologies.

3.2 Event extraction

For event extraction, we applied EventMine<sup>4</sup> a pipeline-based event extraction system using support vector machines (SVM). EventMine takes as input document text and entity annotations, and extracts event structures and modications: EventMine outperforms the best systems participating in the original BioNLP Shared Task 2011 on the GE and ID data sets (with $F$ scores 58.0% and 57.6%, respectively) and is competitive with the best systems on the EPI data set (Kim et al., 2011b; Miwa et al., 2011). Next, we optionally perform a matching step using dictionaries compiled from the UMLS Metathesaurus and OBO resources. We then extract a comprehensive set of features for machine learning, building on orthographic, lexical, syntactic and dictionary match information (see Supplementary information).

EventMine consists of four modules: (i) event trigger detection marks likely triggers and assigns them types, (ii) argument detection identifies likely trigger-argument pairs and assigns them roles, (iii) multi-argument event detection combines trigger-argument pairs into likely event structures and (iv) modification detection assigns modification flags (Negation and...

<sup>3</sup>http://nersuite.nlplab.org

<sup>4</sup>http://www.nactem.ac.uk/EventMine/
3.3 Experimental setup

The annotated data were initially divided into training, development and test sets. The test set was held out during method development and parameter selection. For the final experiment, methods were trained on the combination of training and development data and evaluated on the test set.

We evaluate both entity mention detection and event extraction performance using the standard precision, recall and F score metrics, microaveraged over instance-level true-positive, false-positive and false-negative counts.

For entity mention detection, we apply the evaluation protocol and tools of the BioNLP Shared Task 2011 [Kim et al. 2011] including providing gold entity annotations as given for event extraction. We apply the primary matching criteria defined in the task, which otherwise require event structures to be identical but include the approximate span and approximate recursive relaxations to exact match: the former allows small variation in predicted event trigger spans and the latter permits differences in the secondary arguments of recursive event structures for matches. For detailed definitions, we refer to [Kim et al. 2011].

For event extraction, we adapt the evaluation protocol and tools introduced in the BioNLP Shared Task 2011 [Kim et al. 2011], including providing gold event annotations as given for event extraction. We apply the primary matching criteria defined in the task, which otherwise require event structures to be identical but include the approximate span and approximate recursive relaxations to exact match: the former allows small variation in predicted event trigger spans and the latter permits differences in the secondary arguments of recursive event structures for matches. For detailed definitions, we refer to [Kim et al. 2011].

4 RESULTS AND DISCUSSION

We next present the primary results of the annotation effort and the entity mention detection and event extraction experiments.

4.1 Annotation effort and results

We estimate the concentrated effort to produce the corpus annotation to have totalled approximately 250 hours, of which approximately 100 hours used on guideline development, management and annotation consistency checking. The effort required to produce structured event annotation is thus broadly comparable to the initial effort by domain experts to mark text spans of interest (Wang et al., 2011).

Table 3 presents the overall statistics of the annotated multi-level event extraction (MLEE) corpus. We note that the texts

Table 3. Overall corpus statistics

| Item | Train | Devel | Test | Total |
|------|-------|-------|------|-------|
| Document | 131 | 44 | 87 | 262 |
| Sentence | 1271 | 457 | 880 | 2608 |
| Entity | 4147 | 1431 | 2713 | 8291 |
| Organism | 359 | 126 | 237 | 722 |
| Anatomy | 1844 | 589 | 1166 | 3599 |
| Molecule | 1944 | 716 | 1310 | 3970 |
| Event | 3296 | 1175 | 2206 | 6677 |
| Anatomical | 810 | 269 | 596 | 1675 |
| Molecular | 340 | 125 | 240 | 705 |
| General | 1851 | 627 | 1176 | 3654 |
| Planned | 295 | 154 | 194 | 643 |

See Table 4 for overall and test set statistics.

Table 4. Comparison of corpus statistics with BioNLP Shared Task 2011 corpora annotated using the same representation

| Item | MLEE | EPI | GE | ID |
|------|------|-----|----|----|
| Document | 262 | 1200 | 1224 | 30 |
| Word | 56 | 588 | 253 | 348 | 153 | 153 |
| Entity | 8291 | 15190 | 21616 | 4150 |
| Event | 3296 | 1175 | 2206 | 6677 |
| Anatomical | 810 | 269 | 596 | 1675 |
| Molecular | 340 | 125 | 240 | 705 |
| General | 1851 | 627 | 1176 | 3654 |
| Planned | 295 | 154 | 194 | 643 |

* The ID document count is low as the corpus consists of half-text documents, not abstracts.

Fig. 4. Example Negative regulation (−Reg) event connecting entities at different levels of biological organization

include comparable numbers of molecular and anatomy-level entity mentions, with a lower but still notable number of organism mentions. The event counts show a higher density of anatomical than molecular-level events, although general biological events dominate overall. Overall, 1222 events, or 18% of the total, involve either directly or indirectly (through participating events) arguments at both the molecular and anatomy levels (Fig. 4). Table 3 presents corpus statistics with reference to those for the three largest event-annotated corpora in the recent BioNLP shared task 2011. We note that although the MLEE corpus is smaller than these corpora focusing on the molecular level in terms of e.g. word count, there is less difference in the number of entity annotations, and the MLEE corpus has more event annotations than two of the shared task corpora. The introduced corpus thus has a very high density of event annotations, which we attribute in part to the novel entity and event types allowing a more comprehensive representation of statements in text.

We refer to Supplementary Material Section 1.3 for an evaluation of the corpus annotation consistency.
4.2 Entity mention detection

The overall evaluation results for entity mention detection are listed in Table 5. We find a consistent benefit from the use of the lexical resources, with e.g. a 3.6% point improvement in F score (15% reduction in error) for strict matching. As expected, evaluated performance is notably higher under the relaxed criteria, in particular for right boundary matching. This suggests comparatively many errors in the choice of noun premodifiers included in annotation span, a distinction that may not be of critical importance for many applications.

Table 5. Overall entity mention detection results (prec/rec/F score)

| Model         | Exact | Left boundary | Right boundary |
|---------------|-------|---------------|----------------|
| Base          | 77.03/69.18/72.89 | 79.85/71.72/75.57 | 82.47/74.07/78.04 |
| Dictionary    | 79.49/73.77/76.52 | 82.59/76.64/79.50 | 84.68/78.58/81.52 |

Table 6. Entity mention detection results by category for dictionary model (prec/rec/F score)

| Category          | Exact | Left boundary | Right boundary |
|-------------------|-------|---------------|----------------|
| ANATOMICAL        | 90.82/82.10/86.24 | 91.79/82.97/87.16 | 91.79/82.97/87.16 |
| ORGANISM          | 80.91  | 72.05         | 76.22          |
| MOLECULE          | 90.82  | 72.05         | 76.22          |

4.3 Event extraction

The overall results for event extraction using EvenMine are presented in Table 7. The results demonstrate that the stacked model incorporating information from the previously introduced GE corpus outperforms a purely corpus-internal model. Although the improvement from incorporating the independently annotated out-of-domain data is somewhat modest, the result does indicate that the annotation has met its aim to maintain compatibility with this key resource for molecular-level event annotation.

Table 7. Overall event extraction results

| Model          | Prec | Rec | F score |
|----------------|------|-----|---------|
| Base           | 56.53 | 48.72 | 52.34   |
| Stacking (GE)  | 56.38 | 50.77 | 53.43   |

Table 8. Event extraction results by category for stacked model

| Category          | Prec | Rec | F score |
|-------------------|------|-----|---------|
| ANATOMICAL        | 68.44 | 75.63 | 71.86   |
| MOLECULAR         | 56.68 | 51.96 | 54.22   |
| PLANNED           | 43.87 | 38.99 | 41.29   |
| MODIFICATION      | 47.95 | 29.92 | 36.85   |

Event categories as defined in Table 2. Stacking gives performance for Negation and Speculation detection.

5 CONCLUSION

We have presented the MLEE corpus, a resource aiming to extend the coverage of resources and methods for structured event extraction from the molecular level to encompass all levels from the subcellular to the organism. Experiments using state-of-the-art entity mention detection and event extraction methods demonstrated that the newly proposed extraction targets can be met with reasonable performance using the MLEE corpus, with approximately 80% overall F score for entity mention detection and over 50% F score for event extraction using standard evaluation criteria.

In future work, we will focus on the extension of the annotations and extraction methods to improve the domain independence of...
our annotation to allow the application of the introduced extraction methods at large scale to automatically annotate the entire available literature. The results of these extraction efforts will be made available through search systems such as MEDIE to further improve access to the biomedical literature by facilitating structured semantic queries across multiple levels of biological organization, for example to find statements regarding the inhibition of organ growth by specific molecular-level entities or events.

All resources introduced in this study, including the annotated corpus, guidelines, the evaluation tools and the methods are available from http://nactem.ac.uk/MLIE/

ACKNOWLEDGEMENTS

We thank Xinglong Wang and Iain McKendrick for their advice in interpreting the annotation of the reference corpus and the anonymous reviewers for their many insightful comments.

Funding: This work was funded by UK Biotechnology and Biological Sciences Research Council (BBSRC BB/G013160/1).

Conflict of Interest: none declared.

REFERENCES

Amanatidis, S. et al. (2010) Event extraction for systems biology by text mining the literature. Trends Biotechnol., 28, 381–390.

Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. Nat. Genet., 25, 25–29.

Bida,M. and Hume,L. (2011) Desiderata for ontologies to be used in semantic annotation of biomedical documents. J. Biomed. Inform., 44, 94–101.

Bjorn,J. et al. (2010) Complex event extraction at PubMed scale. Bioinformatics, 26, 1882–1890.

Bredenfelder.O (2004) The unified medical language system (UMLS) integrating biomedical terminology. Nucleic Acids Res., 32 D267-D270.

Brisac,G. et al. (2012) BioNLP 2011 Shared Task—the bacteria track. BMC Bioinformatics, 13, S3.

Brinkman,R. et al. (2010) Modeling biomedical experimental processes with OBI. J. Biomed. Semant., 1(suppl. 5), S7.

Carmel,E. and Jan,R.R. (2000) Angiosarcoma in cancer and other diseases. Nature, 407, 249–257.

Fan,K.-E. et al. (2008) LEBLINEAR: A library for large linear classification. J. Mach. Learn. Res., 9, 1871-1874.

Gemer,M. et al. (2010) LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics, 11, 85.

Haendel,M. et al. (2008) CAID—the common anatomy reference ontology. Anat. Oncol. Bioinformatics, pages 327–349.

Kim,J.-D. et al. (2004) Introduction to the bio-entity recognition task at INLPRA. In Proceedings of INLPR 2004, pp. 70–75.

Kim,J.-D. et al. (2008) Corpus annotation for mining biomedical events from literature. BMC Bioinformatics, 9, 10.

Kim,J.-D. et al. (2011a) Extracting bio-molecular events from literature—the BioNLP'09 shared task. Association for Computational Linguistics, Compu. Intell., 27, 513–540.

Kim,J.-D. et al. (2011b) Overview of BioNLP Shared Task 2011. In Proceedings of the BioNLP 2011 Shared Task, Portland, Oregon, USA.

Kiran,M. (2002) Systems biology: a brief overview. Science, 295, 1662–1664.

Krattinger,M. et al. (2007) Assessment of the second BioCreative PPI task: automatic extraction of protein-protein interactions. In Proceedings of BioCreative II, pp. 41–54.

Kumar,A. et al. (2004) Biomedical informatics and granularity. Comp. Facet. Genomics, 5, 501–508.

Lafferty,J. et al. (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001, Williamstown, MA, USA.

Mayhew,D. et al. (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res., 33(Database issue), D54-D58.

Mirmehdi,M. et al. (2012) Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. Bioinformatics, 28, 1759–1765.

Miyao,Y. et al. (2009) Evaluating contributions of natural language parsers to protein–protein interaction extraction. Bioinformatics, 25, 394–400.

Mungall,C.J. et al. (2011) Cross-product extensions of the gene ontology. J. Biomed. Informatics, 44, 80–86.

Nédélec,C. (2005) Learning language in logic: genetic interaction extraction challenge. In Proceedings of IJL’2005, pp. 31–37.

Oba,T. et al. (2011a) Event extraction for DNA methylation. J. Biomed. Semant., 2(suppl 5), S5.

Oba,T. et al. (2011b) Pathway curator support as an information extraction task. In Proceedings of LBM 2011, Singapore.

Okazaki,N. (2007) CRFsuite: a fast implementation of conditional random fields(CRFs).

Pysalos,S. et al. (2011) Towards exhaustive protein modification event extraction. In Proceedings of BioNLP 2011, Portland, Oregon, USA.

Pysalos,S. et al. (2012a) Annotation guidelines for multi-level event extraction corpus. Technical Report. National Centre for Text Mining, Manchester, UK.

Pysalos,S. et al. (2012b) Learning to classify anatomical entities using open biomedical ontologies. J. Biomed. Semantics. In press.

Pyysalo,S. et al. (2012c) Overview of the IE, EPI and REL tasks of BioNLP Shared Task 2011. BMC Bioinformatics, 13, S2.

Rosse,C. and Mejino,J. (2003) A reference ontology for biomedical informatics: the foundational model of anatomy. J. Biomed. Inform., 36, 476–500.

Rosse,C. and Mejino,J. (2008) The foundational model of anatomy ontology. Anot. Oncol. Bioinformatics, 6, 59–117.

Sage,B. and Tsujii,J. (2007) Dependency parsing and domain adaptation with LR models and parser ensembles. In Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL, Association for Computational Linguistics, Vol. 7, pp. 1044–1050.

Smith,B. (2003) Ontology. In Floridi L.(ed.). The Blackwell Guide to the Philosophies of Computing and Information. Blackwell, pp.155–166.

Smith,B. et al. (2005) On carcinomas and other pathological entities. Comp. Facet. Genomics, 6, 379–387.

Smith,B. et al. (2007) The ORO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol., 25, 1255–1255.

Smits, M. et al. (2012) i2b2: a web-based tool for idp-assisted text annotation. In Proceedings of IJCL, 2012, Association for Computational Linguistics, pp. 102–107.

Tanabe,L. et al. (2005) GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics, 6(suppl 1), S3.

Tsuruoka,Y. and Tsujii,J. (2005) Bidirectional inference with the easiest-first strategy using domain adaptation and coreference resolution. Bioinformatics, 21, S7–S10.

Wang,X. et al. (2011) Automatic extraction of angiogenesis bioprocess from text. Bioinformatics, 27, 2710–2717.

Wilbur,W. et al. (2007) BioCreative 2 gene mention task. In Proceedings of the Second BioCreative Challenge Evaluation, pp. 7–16.

Yeh,A. et al. (2005) BioCreative task 1a: gene mention finding evaluation. BMC Bioinformatics, 6(suppl 1), S2.