Event extraction across multiple levels of biological organization

Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Jun’ichi Tsujii and Sophia Ananiadou

National Centre for Text Mining and School of Computer Science, University of Manchester, Manchester, UK,
Department of Computer Science, University of Tokyo, Tokyo, Japan and 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 entities and event extraction tasks.

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

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 organism-level effects to organism-level outcomes (Fig. 1). 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 (Babcock et al. 2005) and on the extraction of pairwise relations between these representative, for example, protein–protein interactions (Krallinger et al. 2007; Niedeck 2008). 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); event-annotated corpora have been introduced for many extraction targets, including DNA methylation, Ohta et al. (2011b), protein modifications Pyysalo et al. (2011) and the molecular mechanisms of infectious diseases Pyysalo et al. (2012b); event extraction methods have been applied to automatically analyze all 20 million PubMed abstracts Genome et al. (2011); and event extraction analyses are being integrated into literature search systems such as MEDIE1 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 events1. 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

1To whom correspondence should be addressed.

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
We apply the specific event representation first formalized in

We selected as the starting point for our study a recently introduced

levels of biological organization.

texts thus represent a good test case for structured IE across multiple

is increasingly understood (Carmeliet and Jain, 2000), and domain

and other organism-level pathologies and whose molecular basis

corpus of 262 PubMed abstracts on angiogenesis, the development

in the event in a specific role (e.g. Protein

entities or other events—each of which is identified as participating

event structures (or

in numerous resources and methods introduced since. In this

ontological basis of the existing entity annotation.

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 a
tissue/organ-level process that is closely associated with cancer
and other organism-level pathologies and whose molecular basis
is increasingly understood (Carmeliet and Jain, 2000), and domain
texts thus represent a good test case for structured IE across multiple

levels of biological organization.

The corpus texts were previously annotated by Wang (2011)
using a typed-span representation, marking references to molecular
level entities, cells, tissues and domain-relevant processes. We
use these annotations created by domain experts as a reference
for identifying statements of interest for our annotation, which
focuses on introducing structured event annotation and solidifying
the ontological basis of the existing entity annotation.

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.

Given the starting point of the existing corpus annotations,
our event annotation effort proceeds from spans to a structured
representation that can represent complex associations between
arbitrary numbers of entities (Fig 1) and many other aspects that the
typed-span representation cannot, such as the direction of causality
(Fig 2).

In addition to selecting the general form of representation, to
define a specific event annotation scheme, we must also fix
the annotated entity and event types as well as the roles, participant
scopes and modifiers applied. For these, we build on previously
introduced resources targeting the molecular level, basing our
extensions on domain ontologies.

2.3 Ontological basis

We take as basic the division between continuants (or endurants)
and occurred (perdurants, processes or events) (see e.g. Smith, 2003)
and adopt the general principle followed also in major
previously introduced 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) foundry (Smith et al., 2007). In brief, before
primary annotation, we analyze 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 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 (Kim 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 (Rosse and Mejino, 2003, 2008), as outlined in our
previous work on anatomical entities (Pyyusal et al., 2012). To
account for pathological anatomy-level entities (e.g. glioma)—out of

"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.

Although the existing corpus annotation of Wang et al. (2011) identifies
such mentions, they are typed nonspecifically, using e.g. Positive
regulation to mark ‘development’ and Negative regulation for ‘cell
death’.

Fig. 1. Example sentence with event annotation. PROT - REG and CELL
comp. abbreviated for Protein, NEGATIVE regulation and Cell component,
respectively.

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

<table>
<thead>
<tr>
<th>Type</th>
<th>Term(s)</th>
<th>Examples</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Single cell org:organelle, multi-cellular org:organelle</td>
<td>Human, mice, C. albicans</td>
<td>722</td>
</tr>
<tr>
<td>Anatomy</td>
<td>Anatomical system:organ system</td>
<td>Head, thorax, hindlimb, legs</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Organism subdivision:organism subdivision</td>
<td>Central nervous system, pulmonary system</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>Compound organ:organelle</td>
<td>Heart, eyes, skin</td>
<td>176</td>
</tr>
<tr>
<td></td>
<td>Multi-tissue structure:multi-tissue structure</td>
<td>Blood vessel, perivascular membrane, lymph nodes</td>
<td>554</td>
</tr>
<tr>
<td></td>
<td>Portion of tissue:part of tissue</td>
<td>Endothelium, adipose tissue, capillary</td>
<td>426</td>
</tr>
<tr>
<td></td>
<td>Cell:cell</td>
<td>Endothelial cells, HUVECs, pericyte, cancer cells</td>
<td>1198</td>
</tr>
<tr>
<td></td>
<td>Cellular component:cellular component</td>
<td>Nuclei, focal adhesions, extracellular matrix</td>
<td>145</td>
</tr>
<tr>
<td></td>
<td>Developing anatomical structure:developing anatomical structure</td>
<td>Embryo</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Organism substance:organism substance:organ substance</td>
<td>Blood, serum, plasma, urine</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>Immaterial anatomical entity:immaterial anatomical entity</td>
<td>Lumen, preperitoneal space, marrow cavity</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Pathological formation:pathological formation</td>
<td>Tumor, colorectal cancer, gliomas</td>
<td>910</td>
</tr>
<tr>
<td></td>
<td>Drug or compound:drug</td>
<td>Oxygen, ethanol, bromocriptine, thalidomide</td>
<td>944</td>
</tr>
<tr>
<td></td>
<td>Gene or gene product:gene</td>
<td>VEGF, p53, IL-6, endostatin, thrombin</td>
<td>2962</td>
</tr>
</tbody>
</table>

Labels in [ ] identify informal categories used in evaluation.
1 Annotated also in previously introduced event extraction resources. e, identifies a term t in an ontology o; ontology identifiers are OBO Foundry prefixes (namespaces).
2 This annotation strategy can be viewed as partly analogous to efforts to make GO term structure explicit [Mungall et al. 2005].

Fig. 3. Annotation with detailed GO terms (top; hypothetical) and event annotation with general types (bottom; applied)

For event types, we draw primarily on the biological process subontology 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, references to specific processes included in GO through composite terms such as regulation of heart growth, are captured using the explicitly structured representation [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. 2005], 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 2.

For event participants, we apply otherwise standard roles included also in previous efforts (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—canonical anatomy and physiological processes—but texts primarily refer to real-world instances that do not fill these exacting criteria [Itada 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 BRAT rapid annotation tool [Stenetorp et al. 2011].
Table 2. Primary event types, argument roles, related ontology terms and annotation counts

<table>
<thead>
<tr>
<th>Type</th>
<th>Arguments</th>
<th>Terms(s)</th>
<th>Examples</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANATOMICAL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CELL PROLIFERATION</td>
<td>Theme</td>
<td>Cell proliferation$_{30}$</td>
<td>proliferating [ECs], [SMCs] accumulated</td>
<td>133</td>
</tr>
<tr>
<td>DEVELOPMENT</td>
<td>Theme</td>
<td>Developmental process$_{30}$</td>
<td>[skin] development, [stress fiber] formation</td>
<td>316</td>
</tr>
<tr>
<td>GROWTH</td>
<td>Theme</td>
<td>Growth$_{30}$</td>
<td>angiogenesis, neovascularization</td>
<td>855</td>
</tr>
<tr>
<td>DEATH</td>
<td>Theme</td>
<td>Death$_{30}$</td>
<td>growth [of arteries], [tumour] growth</td>
<td>169</td>
</tr>
<tr>
<td>BREAKDOWN</td>
<td>Theme</td>
<td></td>
<td>[connective tissue] necrosis, [cell] apoptosis</td>
<td>97</td>
</tr>
<tr>
<td>REMODELING</td>
<td>Theme</td>
<td>Tissue remodeling$_{30}$</td>
<td>[vascular] remodeling, changes [in membrane]</td>
<td>33</td>
</tr>
<tr>
<td>MOLECULAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SYNTHESIS</td>
<td>Theme</td>
<td>Biosynthetic process$_{30}$</td>
<td>[ATP] synthesis, production [of NOS]</td>
<td>17</td>
</tr>
<tr>
<td>GENE EXPRESSION*</td>
<td>Theme</td>
<td>Gene expression$_{30}$</td>
<td>expression [of VEGF]</td>
<td>435</td>
</tr>
<tr>
<td>TRANSCRIPTION*</td>
<td>Theme</td>
<td>Transcription, DNA-dependent$_{30}$</td>
<td>VEGF expression</td>
<td>37</td>
</tr>
<tr>
<td>CATABOLISM*</td>
<td>Theme</td>
<td>Catabolic process$_{30}$</td>
<td>[p53] breakdown</td>
<td>26</td>
</tr>
<tr>
<td>PHOSPHORYLATION*</td>
<td>Theme, Site</td>
<td>Phosphorylation$_{30}$</td>
<td>phosphorylation [of KDR]</td>
<td>33</td>
</tr>
<tr>
<td>DEPHOSPHORYLATION*</td>
<td>Theme, Site</td>
<td>Dephosphorylation$_{30}$</td>
<td>[M(1-1)] dephosphorylation</td>
<td>6</td>
</tr>
<tr>
<td>LOCALIZATION*</td>
<td>Theme, All/From-To-Loc</td>
<td>Localization$_{30}$</td>
<td>[VEGF] colocalized, [VPF was] secreted</td>
<td>450</td>
</tr>
<tr>
<td>BINDING*</td>
<td>Theme, Site</td>
<td>Binding$<em>{30}$, biological adhesion$</em>{30}$</td>
<td>[cell] adhesion, [GDP-bound [RbHsa]]</td>
<td>184</td>
</tr>
<tr>
<td>REGULATION*</td>
<td>Theme, Cause, Site</td>
<td>Biological regulation$_{30}$</td>
<td>[AMSI] modulation [activation of AP-1]</td>
<td>773</td>
</tr>
<tr>
<td>POSITIVE REGULATION*</td>
<td>Theme, Cause, Site</td>
<td>Pos.regulation of biol.proc$_{30}$</td>
<td>[inulin] stimulates [VEGF expression]</td>
<td>1327</td>
</tr>
<tr>
<td>NEGATIVE REGULATION*</td>
<td>Theme, Cause, Site</td>
<td>Neg.regulation of biol.proc$_{30}$</td>
<td>Inhibition [of NO synthase by L-NAME]</td>
<td>921</td>
</tr>
<tr>
<td>PLANNED</td>
<td>Theme</td>
<td>Planned process$_{30}$</td>
<td>injection [of U-995], [UTPl] administration</td>
<td>643</td>
</tr>
</tbody>
</table>

Notes:
- *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 (Okazaki 2004) of conditional random fields (affinity et al. 2000). 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) (Tanghe et al. 2009), effectively matching the performance of the best available systems for the task.

Following initial sentence splitting and tokenization, we perform lemmatization, POS-tagging and shallow parsing using the GENIA tagger (Eureqa and Tunt) (2006). Next, we optionally perform a matching step using dictionaries compiled from the UMLS Metathesaurus (Bodenreider et al. 2004), Entrez Gene (Maglott et al. 2007) resources. 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 (2010) 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 (Kum et al. 2011, Mora et al. 2012).

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...

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.
We estimate the concentrated effort to produce the corpus annotation. We next present the primary results of the annotation effort and the arguments of recursive event structures for matches. For detailed definitions, event trigger spans and the latter permits differences in the secondary relaxations to exact match: the former allows small variation in predicted 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 Supplementary Material Section 1.3 for an evaluation of the corpus annotation consistency.

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 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 4 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 classes in the classification problems. In addition to training EventMine on the newly introduced corpus, we also introduced a set of generalization rules appropriate to the introduced types. We refer to supplementary documentation and Wang et al. (2011) for further details on EventMine.

Table 3. Overall corpus statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Train</th>
<th>Devel</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document</td>
<td>131</td>
<td>44</td>
<td>87</td>
<td>262</td>
</tr>
<tr>
<td>Sentence</td>
<td>1271</td>
<td>457</td>
<td>880</td>
<td>2608</td>
</tr>
<tr>
<td>Word</td>
<td>27875</td>
<td>9610</td>
<td>19103</td>
<td>56588</td>
</tr>
<tr>
<td>Entity</td>
<td>4147</td>
<td>1431</td>
<td>2713</td>
<td>8291</td>
</tr>
<tr>
<td>Organism</td>
<td>359</td>
<td>126</td>
<td>237</td>
<td>722</td>
</tr>
<tr>
<td>Anatomy</td>
<td>1844</td>
<td>589</td>
<td>1166</td>
<td>3599</td>
</tr>
<tr>
<td>Molecule</td>
<td>1944</td>
<td>716</td>
<td>1310</td>
<td>3970</td>
</tr>
<tr>
<td>Event</td>
<td>3296</td>
<td>1175</td>
<td>2206</td>
<td>6677</td>
</tr>
<tr>
<td>ANATOMICAL</td>
<td>810</td>
<td>269</td>
<td>596</td>
<td>1675</td>
</tr>
<tr>
<td>Molecular</td>
<td>340</td>
<td>125</td>
<td>240</td>
<td>705</td>
</tr>
<tr>
<td>General</td>
<td>1851</td>
<td>627</td>
<td>1176</td>
<td>3654</td>
</tr>
<tr>
<td>Planned</td>
<td>295</td>
<td>154</td>
<td>194</td>
<td>643</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Item</th>
<th>MLEE</th>
<th>EPI</th>
<th>GE</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document</td>
<td>262</td>
<td>1200</td>
<td>1224</td>
<td>30</td>
</tr>
<tr>
<td>Word</td>
<td>56588</td>
<td>253628</td>
<td>348908</td>
<td>153153</td>
</tr>
<tr>
<td>Entity</td>
<td>8291</td>
<td>15190</td>
<td>21616</td>
<td>12740</td>
</tr>
<tr>
<td>Event</td>
<td>6677</td>
<td>3714</td>
<td>24967</td>
<td>4158</td>
</tr>
</tbody>
</table>

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

Fig. 4. Example Negative regulation (-Reg) event connecting entities at different levels of biological organization.
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 6 lists a breakdown of performance by entity category for the dictionary model. The detection of Organism mentions is most reliable despite their sparseness in the data, confirming to previous results indicating this entity class to represent a comparatively easy problem (Fernandez et al. 2010). The detection of mentions of entities of the Anatomy and Molecule categories can be performed at broadly comparable accuracy on this corpus containing balanced numbers of annotations of the two, suggesting that fine-grained anatomical entity detection is no more difficult than established molecular level entity detection tasks.

The overall entity mention detection performance, approaching or exceeding 80% in F score depending on evaluation criteria, is a very promising result given the novelty of the task and its many challenging aspects, most obviously that it involves more than 10 distinct entity types. As points of comparison, the best single system at the well-established single-class BioCreative 2 Gene Mention task achieved an F score of 87.2% under matching criteria that in cases accept more than one specific span as correct (Wilson et al. 2003) and the highest-performing system at the original BioNLP/InLPCA shared task, involving the detection of entities of five different types, achieved an F score of 72.6% under the exact matching criterion (Kim et al. 2004).

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.

As for entity mention detection, performance for the best model, at over 50% F score for event extraction, is very promising for a first experiment on the new task. For reference, the best results in the recent, widely attended BioNLP Shared Task 2011 for the same evaluation criteria were 56.0% F score for the GE task, 53.3% F score for the EPI task and 55.6% F score for the ID task (Table 8: Kim et al. 2011b). Reaching this general level of performance suggests that the task is feasible for current event extraction technology and that the annotation consistency and the size of the introduced corpus are sufficient for reliable extraction.

Table 7 gives a breakdown of the event extraction performance by category. Interestingly, we find that events involving anatomical entities are more reliably extracted than those involving molecular-level ones, despite the model incorporating information from a corpus with a larger number of molecular level event annotations than the total number of annotations in the MLEE corpus. This is a very encouraging finding for event extraction for anatomical processes, indicating that the representation and extraction methods are well suited for the task.

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/IGLEE/.

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

Amita, S. et al. (2010) Event extraction for systems biology by text mining


Bada, M. and Hunter, L. (2011) Desiderata for ontologies to be used in semantic

Learn. Res. 9, 1871–1874.

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

Ontolo. Bioinformatica, pages 327–349.

Chang, Y. et al. (2004) Introduction to the bio-entity recognition task at JNLPIA.

Chen, F. et al. (2008) CORPUS annotation for mining biomedical events from literature.
BMC Bioinformatics. 9, 10.

Chen, J. D. et al. (2011a) Extracting bio-molecular events from literature—the
27, 513–540.

BioNLP 2011 Shared Task, Portland, Oregon, USA.


extraction of protein-protein interactions. In Proceedings of BioCreative II,
pp. 41–54.

Genomics, 5, 501–508.

and labeling sequence data. In Proceedings of ICML 2001, Williamstown, MA,
USA.


Miwa, M. et al. (2012) Boosting automatic event extraction from the literature

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

Informatics. 44, 80–86.


Ohta, T. et al. (2011b) Pathway curation support as an information extraction task. In


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

Pyykko, S. et al. (2012a) Annotation guidelines for multi-level event extraction corpus.

Pyykko, S. et al. (2012b) Learning to classify anatomical entities using open biomedical

Pyysalo, S. et al. (2012c) Overview of the ID, EPI and REL tasks of BioNLP Shared


Ontolo. Bioinformatica, 6, 59–117.

and parse ensembles. In Proceedings of the CoNLL Shared Task Session of EMNLP-


Genomics, 6, 379–387.

Smith, B. et al. (2007) The ORO Foundry: coordinated evolution of ontologies to support

Stemlot, P. et al. (2012) lext: a web-based tool for nlp-assisted text annotation. In
Proceedings of EACL 2012, Association for Computational Linguistics, pp. 102–
107.

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

Tsuruoka, Y. and Tsujii, J. (2005) Bidirectional inference with the easiest-first strategy
for tagging sequence data. In Proceedings of HLT-EMNLP, Association for

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

BioCreative Challenge Evaluation, pp. 7–16.

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