BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 16 2012, pages 2154–2161
doi:10.1093/bioinformatics/bts332

Data and text mining

Advance Access publication June 17, 2012

BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events

Martin Gerner[1,*,†], Farzaneh Sarafraz[2,†], Casey M. Bergman[1] and Goran Nenadic[2]

[1]Faculty of Life Sciences, University of Manchester, Manchester, M13 9PT and [2]School of Computer Science, University of Manchester, Manchester, M13 9PL, UK

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Although the amount of data in biology is rapidly increasing, critical information for understanding biological events like phosphorylation or gene expression remains locked in the biomedical literature. Most current text mining (TM) approaches to extract information about biological events are focused on either limited-scale studies and/or abstracts, with data extracted lacking context and rarely available to support further research.

Results: Here we present BioContext, an integrated TM system which extracts, extends and integrates results from a number of tools performing entity recognition, biomolecular event extraction and contextualization. Application of our system to 10.9 million MEDLINE abstracts and 234 000 open-access full-text articles from PubMed Central yielded over 36 million mentions representing 11.4 million distinct events. Event participants included over 290 000 distinct genes/proteins that are mentioned more than 80 million times and linked where possible to Entrez Gene identifiers. Over a third of events contain contextual information such as the anatomical location of the event occurrence or whether the event is reported as negated or speculative.

Availability: The BioContext pipeline is available for download (under the BSD license) at http://www.biocontext.org, along with the extracted data which is also available for online browsing.

Contact: martin.gerner@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 22, 2011; revised on May 30, 2012; accepted on June 1, 2012

1 INTRODUCTION

The amount of information available in the biomedical literature is increasing rapidly, with over 2000 articles published daily (http://www.nlm.nih.gov/bsd/index_stats_comp.html). While the information available in these articles (now exceeding 18 million in number) represents a vast source of knowledge, its sheer size also presents challenges to researchers in terms of discovering relevant information. Efforts in biomedical text mining (TM) seek to mitigate this problem through systematic extraction of structured data from literature (Lu, 2011). To date, progress in biomedical TM research has primarily focused on tools for entity recognition (locating mentions of species, genes, diseases, etc.) and the extraction of gene/protein relationships (Krallinger *et al.*, 2008a,b).

Recently, there has been increasing interest to develop TM tools for the extraction of information about a wider array of biological and molecular processes (often referred to as 'events'), such as expression, phosphorylation, binding and regulation of genes and proteins. Community challenges (Kim *et al.*, 2009, 2011) have shown that extracting such events is often difficult because of the complex and inconsistent ways in which such processes are reported in the literature (Zhou and He, 2008). In addition, most efforts to extract events have been restricted to limited-scale studies or abstracts. Although some event extraction tools are now publicly available, their usefulness for supporting biological discovery is still unknown given the difficulties in applying and integrating data from these systems on a large scale.

In this article we present BioContext, an integrated TM system which extracts, extends and integrates results from a number of TM tools for entity recognition and event extraction. The system also provides contextual information about extracted events including anatomical association and whether extracted processes have been reported as speculative or negated (i.e. not taking place). In addition to making the integration platform available under an open-source license, we also provide the data resulting from processing the whole MEDLINE and the open-access subset of PubMed Central (PMC) for batch download and online browsing.

2 BACKGROUND

Biomolecular events are frequently reported and discussed in the literature, and are critical for understanding a diversity of biological processes and functions. Although some databases exist that contain information about certain types of molecular events (e.g. protein–protein interactions, PPIs; Ceol *et al.*, 2009; Szklarczyk *et al.*, 2011), extraction and contextualization of a more general set of events using TM systems will present a valuable addition to manually curated data and enable focused navigation of the literature through a variety of biological processes.

Identification of molecular events in the literature has been the topic of several recent text mining challenges (Kim *et al.*, 2009; Krallinger *et al.*, 2008a,b). The shared task 1 of BioNLP'09 (Kim *et al.*, 2009), for example, aimed to identify and characterize nine types of molecular events: *gene expression, transcription, protein catabolism, localization, phosphorylation, binding, regulation, positive regulation* and *negative regulation*. Depending on the event type, the task included the identification of either one (for the

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

first five event types mentioned above) or more (for *binding*) participating proteins/genes (sometimes referred to as themes). Regulatory events could also have a cause (which could be a protein/gene or another event) in addition to one theme/target of regulation (also a protein/gene or another event). The task also included the identification of a textual span (called 'trigger') that indicated the occurrence of an event. For example, the sentence 'MDM2 acts as a negative regulator of p53 expression' contains two events: (i) a 'gene expression' event, with the theme p53, and (ii) a 'negative regulation' event, where the theme is the gene expression event in (i) and the cause is MDM2. The trigger for the first event is the word 'expression' and the trigger of the regulatory event is 'negative regulator'.

Named entity recognition (NER, locating entities in text) is typically performed before information extraction. Entity classes that have received attention vary widely and include genes/proteins (Leaman and Gonzales, 2008; Settles, 2005), species (Gerner *et al.*, 2010a) and chemical molecules (Hawizy *et al.*, 2011). Recognized entities may also be normalized (i.e. linked to standard database identifiers) in order to enable integration of extracted information with other biological data [e.g. linking gene and protein mentions in the literature to Entrez Gene using GNAT (Hakenberg *et al.*, 2011) or GeneTUKit (Huang *et al.*, 2011)].

Despite increased interest and efforts, only a few general biomolecular event extraction tools are publicly available. The Turku event extraction system (TEES; Björne *et al.*, 2009) combines a machine learning approach (relying on dependency parse graph features) with a rule-based post-processing step to identify complex, nested events. EventMine (http://www.nactem.ac.uk/EventMine/), based on the work of Miwa *et al.* (2010) also uses machine-learning methods and a set of rich features. The Stanford Biomedical Event Parser (McClosky *et al.*, 2011), which uses dependency parses to extract events, has also been made available very recently. Finally, a recent publication by Kano *et al.* (2011) describes the creation of a bio-event meta-service, which would make nine different event extractors (including EventMine and TEES) available through U-compare. However, this system is currently not yet available, and it is not clear if sufficient computational resources would be available for it to perform large-scale document processing.

3 MATERIALS AND METHODS

We designed and implemented an integrated TM system, called BioContext, which extracts, extends and integrates mentions of molecular events in the biomedical literature. The following sections describe the system architecture and components (Section 3.1), integration and event expansion methods (Section 3.2) and the evaluation approaches (Section 3.3).

3.1 System overview, architecture and components

Figure 1 shows an overview of our TM system for large-scale integrated extraction of biomolecular events. Processing is performed in four stages: NER, grammatical parsing, event extraction and contextualization. Each stage is composed of several components, which are described in detail in the following sections. In some cases, outputs from multiple components are merged prior to use by other components. To illustrate the main stages, consider the example sentence 'Interleukin 6 is probably not expressed in the spleen'. In this example, 'Interleukin 6' and 'spleen' are first recognized by the gene/protein and anatomical NER components, respectively (Stage 1). The event extractors then use grammatical processing (Stage 2) to identify the internal structure of the input sentence, and to recognize that it discusses the gene expression event involving Interleukin 6 (Stage 3). Finally, in Stage 4, the extracted event is placed into context: the anatomical association component recognizes that the Interleukin 6 expression relates to the spleen, and the negation/speculation component identifies this event as both negated and speculative.

To facilitate the efficient execution of tools and merging of data, we constructed a lightweight TM integration framework, which we call TextPipe. Although other integrative TM frameworks are available [e.g. UIMA and GATE (Cunningham *et al.*, 2011)], we developed TextPipe since we needed a system which was both more lightweight than what was already available and, more importantly, could be easily modified and optimized for any stability or performance problems we encountered (see Section 4.1). TextPipe makes extensive use of modularization, parallel processing, database optimization and error handling/recovery to address various practical challenges when applying many TM tools to large datasets of abstracts and full-text articles. Tools are wrapped as TextPipe components (treated as black boxes internally) by implementing two simple functions: one to specify the output fields, and another to call the main method of the tool. Data are communicated in the form of lists of key-value pairs.

TextPipe components are either applied directly to documents or run as services on demand. They do not need to provide a list of dependencies. Instead, during run-time, they connect directly to other components, providing the document(s) that need to be processed, and fetching the output of those components to use as their input. Computed results can be stored in a database for later re-use to avoid processing of the same document multiple times.

The components that have been implemented and integrated in BioContext are explained as follows.

Stage 1. NER: identification of genes, species and anatomy

In the first stage, identification of gene and protein mentions is performed by GeneTUKit (Huang *et al.*, 2011) and GNAT (Hakenberg *et al.*, 2011; Solt *et al.*, 2010). To the best of our knowledge, these tools are the only tools available that are capable of high-accuracy gene/protein normalization and are practically applicable to large-scale datasets. Whereas GeneTUKit performs normalization for any species, GNAT was limited to 30 of the most commonly studied organisms. Both tools were configured/adapted to utilize BANNER (Leaman and Gonzales, 2008) for gene/protein name recognition in order to reduce the number of false positive (FP) results. In addition, GNAT was modified to also return any entities recognized by BANNER that could not be normalized. While data extracted using non-normalized entities will have more limited use, they should reduce the number of errors that the event-extraction systems make due to incomplete gene/protein information.

The outputs from both gene/protein NER systems are merged. We used a confidence level cut-off of 0.01 for GeneTUKit, and all results from GNAT. If the two tools have identified overlapping spans, then we create a new span with the union of their coordinates. If the tools have assigned different Entrez

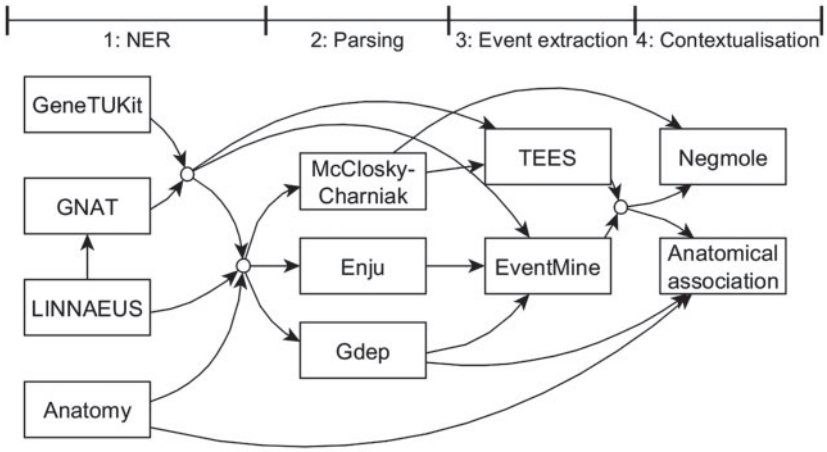

Fig. 1. System architecture. Each box represents the application of one tool (see main text for description), and each arrow represents the transfer of data from one tool to another. Circles represent data merging and post-processing. Entity recognition is performed by GeneTUKit (genes), GNAT (genes), LINNAEUS (species) and GETM's anatomical NER component. Parsing is performed by the McClosky–Charniak, Enju and Genia dependency parsers. Event extraction is performed by TEES and EventMine. In the final contextualization step, Negmole detects whether events are negated and/or speculative, and events are associated to anatomical entity mentions. Additional document parsing functions provide input to the system and database storage functions handle outputs (but are not shown here)

Gene identifiers in the original overlapped spans, then priority is given to the GeneTUKit normalization, as it was ranked higher in BioCreative III gene/protein normalization challenge than GNAT.

Identification of species mentioned was performed using LINNAEUS (Gerner *et al.* 2010a), and recognition of anatomical locations (e.g. brain, T cells) and cell-lines (acting as proxies for anatomical locations, e.g. HeLa for cervical cells) was performed by the anatomical NER system from GETM (Gerner *et al.*, 2010b).

Stage 2. Grammatical parsing

In the second stage, a number of grammatical parsers process the text to determine the structure of the processed sentences. Parsing was done by the McClosky–Charniak constituency parser (McClosky *et al.*, 2006), the Enju constituency parser (Sagae *et al.*, 2008) and the Gdep dependency parser (Sagae and Tsujii, 2007), as these are requested by the down-stream modules (Stages 3 and 4). To increase the accuracy of the parsers when applied to sentences with long and complex entity names, we performed 'semantic tokenization' by ensuring that multi-word phrases that are identified as entity names (e.g. 'acid phosphatase' or 'T cells') were treated as single tokens.

Stage 3. Event extraction

For identification of event mentions, we used TEES and EventMine. Similarly to gene NER tools, these were chosen since they were the only tools available at the time with large-scale processing capabilities and reasonable performance. Both systems recognize events from the nine BioNLP types (gene expression, transcription, protein catabolism, localization, phosphorylation, binding, regulation, positive regulation and negative regulation) and use gene/protein NER results from the first stage. The output from each system consists of information about the event type, trigger and participants. TEES relies on output from the McClosky–Charniak parser, whereas EventMine uses results from the Enju and Gdep parsers. The data extracted by TEES have previously been released for the 2009 MEDLINE baseline release (Björne *et al.*, 2010a,b). We utilized this data for any documents that were in the 2009 MEDLINE baseline release by mapping the entities in the TEES data to those extracted by us in Stage 1 using positional overlaps. Events that referred to entities that could not be mapped to our entities were not included. Further, additional TEES extractions were performed for the remaining documents (the full-text PMC documents and 1 412 095 additional abstracts that only are available in the 2011 MEDLINE baseline release). We performed event extraction with EventMine for all documents.

Before contextualization (Stage 4), the outputs of the two event extraction systems are post-processed, expanded and integrated as described in Section 3.2.

Stage 4. Contextualization

Events can occur in different places and under different conditions, and this information is key for the comprehensive understanding of the biomedical processes. In the fourth stage, extracted events are enriched with contextual information, including species involved, anatomical locations associated with the event and whether extracted events have been reported as speculative or negated. Anatomical locations are linked to events using an expanded version of the GETM method described in Gerner *et al.* (2010b), which relies on Gdep dependency trees to link events and associated anatomical entities. Speculative and negated events are recognized by an extended version of the negation-detection system Negmole (Sarafraz and Nenadic, 2010). Negmole uses machine learning and relies on constituency parse trees from the McClosky–Charniak parser.

3.2 Event integration and post-processing

Events extracted from TEES and EventMine are compared to determine if they refer to the same mention. Two events extracted from a given sentence are considered to be the same if their type and participants match (using approximate span matching, allowing overlap). If the event is 'nested' (e.g. regulates another event), the comparison is performed recursively. Note that we do not require the triggers to match, as they do not convey any additional biological information.

The results of event extraction are additionally post-processed to improve precision (by eliminating likely FPs) and coverage (by inferring additional events), as follows:

(i) *Removing probable FPs.* After studying a sample of the merged output from the large-scale MEDLINE processing, we noted certain patterns that contributed towards clearly incorrectly extracted events. We therefore designed post-processing methods that discriminated against likely FPs, similarly to our previous study (Sarafraz *et al.*, 2009). We have not observed the patterns removing any events that should not have been removed. The rules were based on (a) event chains (sequences of 'nested' events linked through regulation); and (b) event triggers (keywords indicating the presence of an event) as follows.

(a) *Negative discrimination based on the event chains*: For nested regulation events, one or more of the participants can be other events. Common likely FPs were events that are circularly nested (e.g. E1 causes E2, and E2 causes E1) or where there is a long, potentially indefinite chain of events (i.e. E1 causing E2 causing E3 and so on). For example, in one instance, TEES found a chain of 211 769 connected events. We noticed that there were very few instances in the BioNLP'09 training data where events are nested further than two levels, and there are no circularly nested events. Therefore, all events with a 'nestedness' level above 2 were removed.

(b) *Negative discrimination based on the event trigger*: Events characterized with unlikely triggers are also removed, for example, events with very short triggers (one or two characters, mostly consisting of punctuation, single letters or abbreviations). We compiled a whitelist of 11 short words (e.g. 'by') that *could* be triggers, and a blacklist of 15 longer words (common English stop words) that were often recognized incorrectly as event triggers. Events that had a trigger from the whitelist were not removed, whereas events that had a trigger from the blacklist were removed. In addition, capitalized triggers that did not occur in the very beginning of the sentence were also removed as they were likely to be proper nouns. For example, an event with the (incorrect) trigger 'Region' from the sentence 'The prevalence of urinary lithiasis in children in Van Region, Turkey.' (PMID 20027811) would be removed (here, the incorrect event was 'expression of Van', also representing a gene/protein NER FP).

(ii) *Inferring additional events from enumerated entity mentions.* A number of gene/protein and anatomical entity mentions in MEDLINE (see Section 4) are part of entity 'enumerations', i.e. lists of more than one entity connected within a conjunctive phrase. Event extractors, however, typically ignore enumerations. We hypothesized that, where an event is associated with a gene/protein or anatomical entity that is part of such an enumeration, we could infer additional events by substituting the original entity with each of the other entities in the enumeration. For example, consider the sentence 'In the present study, we describe three novel genes, Dorsocross1, Dorsocross2 and Dorsocross3, which are expressed downstream of Dpp in the presumptive and definitive amnioserosa, dorsal ectoderm and dorsal mesoderm.' (PMID 12783790). Here, gene expression events should be extracted for all three Dorsocross genes, and each of those events should be associated with each of the three anatomical locations mentioned. If any of these nine events are not extracted directly, the enumeration processing would allow them to be inferred indirectly.

To implement this inference, we used regular expression patterns to detect groups of enumerated entities. Where at least one of these entities (e.g. T1) were part of an event (e.g. E1), we constructed a new event E2 with the entity T2, where T2 was mentioned in the same enumeration group as T1. Except for T1, all other properties of E1 were duplicated in E2.

3.3 Evaluation approaches

To measure the impact that different processing steps have on the data as it moves through the pipeline, the performance of different components was evaluated individually, with evaluations of the final components also showing the accuracy of the system as a whole. The gene/protein NER, event extraction and negation/speculation detection components were evaluated against a corpus based on the BioNLP'09 and GENIA corpora (described in the following paragraph). The anatomical associations and the event inference components were evaluated by manual inspection.

The public portion of the BioNLP'09 corpus (Kim *et al.*, 2009), consisting of 800 training documents and 150 development test documents was created from a subset of the GENIA event corpus (Kim *et al.*, 2008) with extensive manual modifications (Kim *et al.*, 2009; Ohta *et al.*, 2009). Only the GENIA entities considered to be genes or gene products were included in the BioNLP'09 corpus. However, many events in the GENIA corpus contain links to protein complexes, which were not included in the BioNLP corpus. For example, many mentions of NF-kappa B that refer to protein complexes were removed from the BioNLP'09 corpus. Because of the importance of protein complexes for biomolecular processes, we decided to expand our definition of event participants to also include protein complexes. Therefore, the BioNLP'09 and GENIA corpora were merged into a new corpus, which we refer to as the B+G corpus. More specifically, the BioNLP'09 corpus was expanded with mentions of entities from the Protein_complex and Protein_molecule GENIA classes. Protein_molecule entities were added since protein complexes often were annotated as Protein_molecule entities in the GENIA corpus (for example, we estimate that NF-kappa B was annotated as Protein_molecule in 38% of cases). We also included any events involving these entities. By merging the two corpora, we could retain the modifications made to the BioNLP'09 corpus but gain additional events that involve protein complexes from GENIA. The merged corpus is used as a gold standard for evaluation of both NER and event extraction, and is available in Supplementary Material S1. In the following, we include protein complexes together with genes and proteins when we refer to gene(s) or protein(s).

When evaluating event extraction components, an extracted event was considered to be a true positive (TP) if all of the following criteria hold: (i) the extracted event type is the same as the event type annotated in the gold standard; (ii) the entity participants are all TPs, and approximately match boundaries with the participants in the gold standard; (iii) the participant types match (theme or cause); and (iv) if any of the participants is an event, it is also a TP, defined recursively.

To the best of our knowledge, no corpus currently exists that provides cross-species normalized gene/protein mentions—all available corpora provide gene/protein annotations at the recognition level (i.e. gene/protein mentions are not normalized to database identifiers). Therefore we can only evaluate the combined data of GNAT and GeneTUKit on the recognition level, and refer to the

Table 1. The total number of gene mentions and the number of normalized, distinct genes recognized by GNAT and GeneTUKit in MEDLINE and PMC

Source	Entity mentions			Distinct entities		
	MEDLINE	PMC	MEDLINE + PMC	MEDLINE	PMC	MEDLINE + PMC
GNAT	35 910 779	12 729 471	48 050 830	227 809	129 244	253 929
GeneTUKit	47 989 353	19 217 778	66 431 789	258 765	143 706	287 218
Intersection	26 281 266	8 638 823	34 479 547	224 604	125 763	249 932
Union	57 618 866	23 308 426	80 003 072	261 412	146 552	290 557

Table 2. The total number of event mentions and the number of distinct events extracted by TEES and EventMine from MEDLINE and PMC

Source	Event mentions			Distinct events		
	MEDLINE	PMC	MEDLINE + PMC	MEDLINE	PMC	MEDLINE + PMC
TEES	19 406 453	4 719 648	23 856 554	6 570 824	1 804 846	7 797 604
EventMine	18 988 271	4 010 945	22 737 258	6 502 371	1 588 178	7 539 364
Intersection	9 243 903	1 331 456	10 455 678	3 080 900	573 903	3 424 372
Union	29 150 821	7 399 137	36 138 134	9 635 566	2 676 257	11 442 462

original papers (Hakenberg *et al.*, 2008, 2011; Huang *et al.*, 2011) for evaluation in terms of normalization.

Likewise, no gold-standard data exist that could be used to evaluate anatomical associations or event inference, so we manually inspected a randomly selected set of 100 extracted events for each component to estimate their levels of precision.

4 RESULTS AND DISCUSSION

We performed two types of experiments to assess the benefits of the data produced by our system: a large-scale data generation experiment to quantify and characterize application of the system on MEDLINE and PMC, and a smaller-scale evaluation of the quality of the data.

4.1 Large-scale application to MEDLINE and PMC

We applied our system to MEDLINE (2011 baseline files, containing 10 946 774 abstracts) and to the open-access subset of PMC (downloaded May 2011, containing 234 117 full-text articles.)

Table 1 shows the number of gene/protein entities (both mentions and distinct entities) extracted from MEDLINE and PMC with the two gene normalization tools. Of the 80 003 072 extracted gene mentions in the union set, 10 261 208 (12.8%) were not normalized, all coming from GNAT. The GeneTUKit and intersection data contain only normalized entities linked to Entrez Gene. We note that only 43% (34 479 547/80 003 072) of all mentions in the MEDLINE+PMC union set were recognized by both NER tools. This is even more extreme in the case of full-text articles (only 37% of all mentions recognized by both tools).

Of the 80 003 072 gene/protein mentions in the MEDLINE and PMC union sets, 11 317 242 (14%) were part of enumerated groups as detected by our patterns. Likewise, of the 56 659 248 anatomical mentions found, 3 489 723 (6.2%) were found to be enumerated. These results suggest that authors study multiple genes/proteins more frequently than they study multiple anatomical locations.

Table 2 presents the number of events extracted from MEDLINE and PMC. The relative volumes of different event types are available in Supplementary Material S2. To estimate the number of distinct events reported, we define two events to be the same if the following are true: (i) they are of the same type; (ii) they involve the same normalized gene entities; or, if non-normalized genes are involved, the gene mention strings match; or, if more than one entity is involved, all pairs match; (iii) either no anatomical entity is associated with neither of the two events; or, if one event is associated with an anatomical entity, the other event should also be associated with an entity normalized to the same anatomical location; (iv) they are both affirmative or both negated; (v) they are both certain or both speculative; (vi) if any of the participants of the events is another event, those nested events also match recursively. We note that in total, almost 11.5 million distinct events could be extracted from the MEDLINE+PMC union set. Of the union of distinct events, only 32% and 21% were recognized by both tools in abstracts and full-text articles, respectively. Similar observations hold for event mentions (32% abstracts, 18% full text), demonstrating complementarity of the event extractors.

Of the 36 138 134 event mentions in the MEDLINE+PMC union set, 1 052 541 (2.9%) were created through the event inference method. Although the percentage of events inferred is low, the absolute number is still large enough to demonstrate its utility.

In terms of contextualization, 13 564 939 events (37.5%) could be associated with an anatomical entity, 1 487 502 (4.1%) were negated and 1 253 133 (3.5%) were speculative. We note that the negation/speculation ratios are slightly lower than those of the combined BioNLP'09 training and development sets (at 6.8% and 5.3%, respectively).

Compared with the previously released dataset of 19.2 million total event mentions extracted from the 2009 release MEDLINE by TEES (Björne *et al.*, 2010a,b), the dataset described here provides additional value in a number of ways, including the addition of nearly 1.5 million MEDLINE abstracts and more than 234 000

full-text PMC articles, normalization of genes and proteins to species-specific identifiers, and association to anatomical locations. In addition, the use of multiple tools for the more challenging aspects (gene/protein NER and event extraction), allows users to query and interpret data depending on whether it was extracted by one or more components of the system.

When scaled to the total number of documents processed, we find an average of 2.7 extracted event mentions per abstract and 31.6 event mentions per full-text article in the union data. Thus, only ∼8.4% of the events stated by authors are in the abstracts. This is similar to results from a previous study, which after manual annotation of 29 full-text articles reported that only 7.8% of claims were made in the abstracts (Blake, 2010). Although events stated in the abstracts can be expected to be more important in general than those stated elsewhere in the article, it still highlights the importance of processing full-text documents. It is therefore unfortunate that only ∼2% of MEDLINE entries have open-access full-text articles that are available for text mining. If the open-access subset of PMC is an indicator of the richness of full-text articles in general, we extrapolate that roughly 300–400 million (31.6 × 10.9 million) further event mentions are described in the full text of articles in MEDLINE.

In addition to the 80.0 million gene/protein mentions and the 36.1 million event mentions, the process of extracting events from MEDLINE and PMC also produced large volumes of other intermediary data that are available and should prove useful to the text mining and bioinformatics communities. This data include 70.9 million species entity mentions, 56.6 million anatomical entity mentions and 133 million parsed sentences from each of the Gdep, Enju and McClosky–Charniak parsers.

We note that processing on the scale reported here presents several challenges. The large number of documents resulted in large computational time requirements. Even using 100 processor cores (in a cluster), a full run of the system required 2–3 months. Testing requirements and tool crashes resulted in the total processing requirements of roughly double that. For example, although the documents in MEDLINE and PMC are generally well-structured, there were outliers that introduced significant problems. Examples we have found included documents over 300 pages long (causing some tools to crash when running out of memory, and others never to finish) and documents that typically confuse TM tools by containing non-ASCII characters or programming or TeX source code (causing every single grammatical parser to crash). We have, therefore, implemented robust general error detection and recovery methods within our framework (TextPipe) to help with unusual processing time, frequent crashes and other external problems, such as network connection timeouts or machine failures.

4.2 Gold-standard evaluation

Evaluation results for the gene/protein NER systems on the 3 000 annotated gene/protein mentions in the B+G corpus are shown in Table 3. Both precision (at 72–80% for the individual systems, with a maximum of 83% for the intersection set) and recall (at 79–84%, with a maximum of 92% for the union set) are similar to what has previously been reported for common recognition tools [BANNER: 85% precision, 79% recall; ABNER: 83% precision, 74% recall; (Leaman and Gonzales, 2008)]. A brief manual inspection of FP and false negative (FN) errors indicate that some of the more common categories of errors include incorrect dictionary matches (e.g. non-gene acronyms matching synonym entries in gene dictionaries), the use of terms by authors that are not in dictionaries, and incomplete manual annotations of the corpora.

Table 3. Gene/protein NER evaluation results

	Precision (%)	Recall (%)	F-score (%)
GNAT	79.8	83.7	81.7
GeneTUKit	72.2	79.1	75.5
Intersection	82.8	70.4	76.1
Union	71.4	92.0	80.4

Table 4. Event extraction evaluation results on the B+G corpus

	Precision (%)	Recall (%)	F-score (%)
TEES	50.4	53.6	51.9
EventMine	45.7	45.5	45.6
Intersection	66.2	36.6	47.1
Union	41.3	62.0	49.6

Table 5. Event extraction results on the B+G corpus, including negation/speculation detection as processed by Negmole

	Precision (%)	Recall (%)	F-score (%)
Intersection	62.6	34.6	44.6
Union	38.8	58.3	46.6

Evaluation of the two event extractors on the 2 607 annotated events in the B+G corpus (Table 4) shows the best precision of 66% (for intersection) and the best recall of 62% (for union). Event type-specific evaluation results are available in Supplementary Material S3. TEES alone provides the best balance between precision and recall (52%). Manual inspection reveals that many FP and FN errors by the event extractors were due to incorrect entity recognition that propagated to the event extraction stage, sentences that were particularly linguistically or semantically complex, and incomplete manual annotation of the corpora.

We note that our evaluation results differ from the previously reported level of 64% precision in (Björne *et al.*, 2010a,b); recall was not reported. However, the evaluation methods between these two studies are different, hindering direct comparisons. In the evaluation of (Björne *et al.*, 2010a,b), 100 events were selected randomly for manual verification, rather than being compared with an already annotated gold-standard corpus. Furthermore, a more inclusive definition of 'entity' was also used, with 'cells, cellular components or molecules involved in biochemical interactions' considered as TPs if recognized.

Evaluation of the event extraction results after performing negation and speculation detection by Negmole is shown in Table 5. In addition to having a correctly extracted event (evaluated in Table 4), events were also required to have both their negation and speculation status correctly identified to be classified as a TP. Only

relatively small differences in data quality can be observed before and after the application of Negmole (relative to the results in Table 4). This is expected since only a small subset of events are negated and/or speculative (3.3% and 4.3% of events in the test corpus were found to be negated and speculative, respectively).

To evaluate the performance of the anatomical association and event inference, we randomly selected 100 events associated with anatomical entities and 100 events produced by the event-inference rules. The events were otherwise selected randomly from the complete union of events in MEDLINE and PMC. Manual inspection of the 100 events associated to anatomical entities showed a precision of 34%. The lack of a gold-standard corpus of anatomical associations prevented estimation of recall for anatomical association. The 100 inferred events showed a precision level of 44%.

Overall, we observed that many FPs and FNs that occurred in the NER stage are propagated to the event extraction stage, and additional FPs and FNs introduced there are in turn propagated to the context association stage. This means that even relatively small error rates can have a large impact on the final results, especially if they occur early in the pipeline.

Finally, we note that evaluations performed using the B+G corpus are limited by the fact that it was drawn by the corpus creators from the set of MEDLINE abstracts containing the MeSH terms 'humans', 'blood cells' and 'transcription factors' and therefore may not be completely representative for MEDLINE as a whole.

5 CONCLUSION

In this article we present an integrated TM system, BioContext and the data produced by it when applied to 10.9 million abstracts in MEDLINE and 234 000 full-text articles in the open-access subset of PMC. The data contain 36.1 million event mentions, representing 11.4 million distinct events describing processes such as gene expression, transcription, catabolism, localization, phosphorylation, binding and regulation of genes and proteins. Over a million of additional event mentions were created through enumerated entity mentions. Event participants are whenever possible linked to Entrez Gene identifiers. The data contain contextual information regarding the associated anatomical locations and whether events are reported as negated or speculative.

The process of generating and integrating this huge volume of data proved challenging and differences observed between the output of individual tools indicate that the identification of events in text is not always easily reproducible on a large scale. Nevertheless, differences among tools in recognized entity and event mentions can be useful when deciding the balance between precision and recall for different applications. As expected, we find that for both gene/protein NER and event extraction, the intersection of multiple tools shifts the balance towards increased precision while the union favours increased recall.

We have made the data available at www.biocontext.org both for batch download and for browsing through a web search interface, giving biologists not only more comprehensive access to data in the literature, but allowing bioinformaticians to run more powerful integrative analyses using information extracted from literature. We also provide the entire set of intermediary data files as well as the integration framework, TextPipe, which can be used either for completely new TM projects or to construct and deploy modified versions of the system described here.

Several tasks remain for future work. Data interfaces could be enhanced by allowing restrictions on which documents results are returned for the search interface, or by importing and integrating the BioContext data into a pubmed2ensembl-style biological data mining portal (Baran *et al.* 2011). We would also like to improve the accuracy through improved filtering techniques. Further, as protein complexes currently are not linked to any reference source, we would like to enable normalization of these entities, either to their constitutive genes or directly to protein complex databases. Finally, a more detailed analysis of the extracted data is also warranted, which will hopefully shed new light on biomolecular events at an unprecedented scale of global understanding.

ACKNOWLEDGEMENTS

We thank Makoto Miwa (the National Centre for Text Mining, University of Manchester) for advance access and help with the EventMine software, Jari Björne (University of Turku) for help with the TEES system, and Nick Gresham (University of Manchester) for technical support.

Funding: The University of Manchester and a BBSRC CASE studentship (to M.G.), BBSRC grant BB/E012868/1 (to C.M.B.)

Conflict of Interest: none declared.

REFERENCES

Baran,J. *et al.* (2011) pubmed2ensembl: a resource for mining the biological literature on genes. *PLoS ONE*, **6**, e24716.

Björne,J. *et al.* (2009) Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the Workshop on BioNLP: Shared Task.* Boulder, Colorado, pp. 10–18.

Björne, J. *et al.* (2010a) Complex event extraction at PubMed scale, Bioinformatics, 26, i382-390.

Björne, J. *et al.* (2010b) Scaling up Biomedical Event Extraction to the Entire PubMed. *BioNLP 2010*, Uppsala, Sweden, pp. 28–36.

Blake,C. (2010) Beyond genes, proteins, and abstracts: identifying scientific claims from full-text biomedical articles. *J. Biomed. Inform.*, **43**, 173–189.

Ceol,A. *et al.* (2009) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.

Cunningham,H. *et al.* (2011) *Processing with GATE*. Department of Computer Science, University of Sheffield.

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

Gerner,M. *et al.* (2010b) An exploration of mining gene expression mentions and their anatomical locations from biomedical text. In *Proceedings of the BioNLP workshop*. Uppsala, Sweden, pp. 72–80.

Hakenberg,J. *et al.* (2008) Inter-species normalization of gene mentions with GNAT. *Bioinformatics*, **24**, i126–i132.

Hakenberg,J. *et al.* (2011) The GNAT library for local and remote gene mention normalization. *Bioinformatics*, **27**, 2769–2771.

Hawizy,L. *et al.* (2011) ChemicalTagger: a tool for semantic text-mining in chemistry. *J. Cheminform.*, **3**, 17.

Huang,M. *et al.* (2011) GeneTUKit: a software for document-level gene normalization. *Bioinformatics*, **27**, 1032–1033.

Kano,Y. *et al.* (2011) U-Compare bio-event meta-service: compatible BioNLP event extraction services. *BMC Bioinformatics*, **12**, 481.

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

Kim,J.D. *et al.* (2009) Overview of BioNLP'09 shared task on event extraction. In *Proceedings of the Workshop on BioNLP: Shared Task*. ACL, Boulder, Colorado, pp. 1–9.

Kim,J.-D. *et al.* (2011) Overview of Genia event task in BioNLP Shared Task 2011. In *BioNLP Shared Task 2011*, Portland, Oregon, USA, pp. 1–6.

Krallinger,M. *et al.* (2008a) Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biol.*, **9** (Suppl. 2), S4.

Krallinger,M. *et al.* (2008b) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.*, **9** (Suppl. 2), S8.

Leaman,R. and Gonzales,G. (2008) BANNER: an executable survey of advances in biomedical named entity recognition. In *Pacific Symp. on Biocomputing*. Hawaii, pp. 652–663.

Lu,Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, **2011**, baq036.

McClosky,D. *et al.* (2006) Effective self-training for parsing. In *HLT-NAACL*, Brooklyn, New York, pp. 152–159.

McClosky,D. *et al.* (2011) Event extraction as dependency parsing. In *Association for Computational Linguistics - Human Language Technologies 2011 Conference (ACL-HLT 2011)*, Portland, Oregon, pp. 1626–1635.

Miwa,M. *et al.* (2010) Evaluating dependency representation for event extraction. In *The 23rd International Conference on Computational Linguistics (COLING 2010)*. pp. 779–787.

Ohta,T. *et al.* (2009) Incorporating GENETAG-style annotation to GENIA corpus. In *BioNLP Workshop*. Boulder, Colorado, pp. 106–107.

Sagae,K. and Tsujii,J. (2007) Dependency parsing and domain adaptation with LR models and parser ensembles. In *CoNLL 2007 Shared Task. Joint Conferences on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*. Prague, Czech Republic, pp. 1044–1050.

Sagae,K. *et al.* (2008) Comparative parser performance analysis across grammar frameworks through automatic tree conversion using synchronous grammars. In *COLING 2008*, Manchester, UK, pp. 545–552.

Sarafraz,F. *et al.* (2009) Biomedical event detection using rules, conditional random fields and parse tree distances. In *BioNLP Workshop*. Boulder, Colorado, pp. 115–118.

Sarafraz,F. and Nenadic,G. (2010) Using SVMs with the command relation features to identify negated events in biomedical literature. *The Workshop on Negation and Speculation in Natural Language Processing*. Uppsala, Sweden.

Settles,B. (2005) ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, **21**, 3191–3192.

Solt,I. *et al.* (2010) Gene mention normalization in full texts using GNAT and LINNAEUS. In *Proceedings of the BioCreative III Workshop*. Bethesda, USA, pp. 137–142.

Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.

Zhou,D. and He,Y. (2008) Extracting interactions between proteins from the literature. *J. Biomed. Inform.*, **41**, 393–407.