Event Interpretation: A Step towards Event-Centred Text Mining

Raheel Nawaz
School of Computer Science
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3091
nawazr@cs.man.ac.uk

Paul Thompson
National Centre for Text Mining
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3091
paul.thompson@manchester.ac.uk

Sophia Ananiadou
National Centre for Text Mining
University of Manchester
131 Princess Street, Manchester
M1 7DN, UK
+44 (0)161 306 3092
sophia.ananiadou@manchester.ac.uk

ABSTRACT
Event-centred text mining facilitates semantic querying of document content, providing greater descriptive power and more focused results than traditional keyword searches. In the biomedical domain, automatic assignment of high-level interpretative information to events, e.g., general information content and level of certainty, is useful for a number of tasks. In this paper we motivate the need for correct interpretation of events and describe a new approach for tackling the problem in the biomedical domain.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis, language parsing and analysis*; J.3 [**Computer Applications**]**:** Life and Medical Sciences – *biology and genetics*.

General Terms
Design, Experimentation, Standardization

Keywords
Bio-event, annotation, event interpretation, meta-knowledge

1. INTRODUCTION
Event-based text mining approaches constitute a promising alternative to the traditional approaches, mainly based on the bag-of-words principle. Events are template-like, structured representations of pieces of knowledge contained within documents. Our work focuses specifically on bio-events, which are dynamic relations within the biomedical domain. Text mining systems that are able to extract such events automatically can allow much more precise and focussed searches than the traditional keyword-based systems. Event-based searches specify one or more constraints on the events to be retrieved, which are not dependent on the precise wording in the text. These constraints could be in terms of the type of the event (e.g., positive regulation) and/or its participants (e.g., the instigator of the event must be a protein).

Although event-based searching can retrieve many more relevant documents than is possible using traditional keyword searches, they typically do not take into account the *interpretation* of the event. For example, a particular event may represent generally accepted knowledge, experimental observations, hypotheses or analyses of experimental results. For the two latter types of event, the author may express varying degrees of certainty regarding the analysis performed. We term these types of interpretative information *meta-knowledge.*

Without access to meta-knowledge, a large number of extracted bio-events will be treated identically by text mining systems, even though their intended interpretations may vary significantly [5, 9]. This would pose a serious problem for users of the system whose information requirements include the ability to distinguish between certain interpretations. For example, a biologist who wishes to update either an incomplete model of a biological process (e.g., a molecular pathway) [6] or a curated biological database [1] would wish to locate only newly-reported, reliable experimental knowledge. Thus, he would be interested only in experimental observations or confident analyses of results, but not in hypotheses or more tentative analyses.

The work reported here describes a novel annotation scheme that can be applied to bio-events to make explicit the meta-knowledge associated with them. The annotation caters for several different types (or *dimensions*) of meta-knowledge that could be specified about an event. The aim of the annotation is to facilitate the training of text mining systems that can extract automatically not only events and their participants but also meta-knowledge associated with the event.

2. EVENT-CENTRED TEXT MINING
The knowledge expressed by events is normally organised around a particular word (the event trigger), which is typically a verb or noun. Each event has one or more participants which describe different aspects of the event, e.g., what causes the event, what is affected by it, where it took place, etc. Participants can correspond to entities, concepts or other events, and are often labelled with semantic roles such as CAUSE, THEME or LOCATION to aid in their interpretation and to facilitate more precise searching

Typically, bio-events themselves, as well as bio-entities that constitute the event participants, are assigned types/classes from an appropriate taxonomy or ontology (e.g., [1]). Figure 1 illustrates a simple sentence, together with a typical template-style representation of the bio-events contained within it.

Queries for relevant events can be carried out through partial completion of a template that specifies constraints regarding the events to be retrieved, in terms of one or more of the following:

1st AMICUS Workshop, October, 2010, Vienna, Austria.

- ontological classes of events e.g. *POSITIVE_REGULATION*.
- specifications of the participants that should be present in the event (in terms of semantic roles).
- restrictions on the values of particular participants, in terms of either actual entities (e.g. *NF-kappa B*) or ontological classes (e.g. *PROTEIN*).

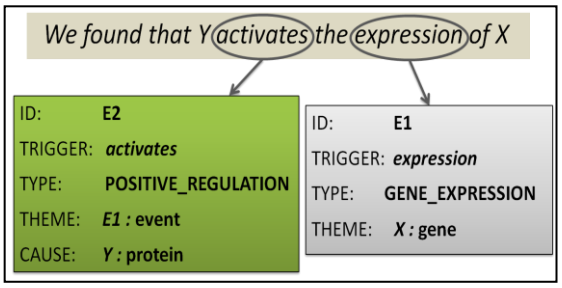

Figure 1. Bio-Event Representation

Searches over events can be more or less specific, depending on the number and nature of the constraints specified.

Event extraction systems are typically trained on collections of texts (corpora) in which events and their participants have been manually annotated by domain experts. Examples include the GENIA Event Corpus [3] and GREC [7]. These corpora allow text mining systems to be trained to recognise and extract events from biomedical texts.

3. INTERPRETATION OF BIO-EVENTS

Existing event annotated corpora within the biomedical domain contain few, if any, annotations that relate to their interpretation. Although more extensive interpretation-focussed annotation has been carried out within the domain at either the sentence level (e.g., [8]) or sentence-fragment level (e.g., [10]), these annotations cannot be used straightforwardly to assign interpretations to bio-events. Often, a sentence will contain several bio-events (e.g. both an experimental method *and* the results of applying this method), each of which has a different interpretation. If an expression of speculation is present (e.g. the word *might*), this may affect only certain events in a sentence.

Our work aims to address this situation through the development of a multi-dimensional annotation scheme that is especially tailored to bio-events. The scheme is intended to be general enough to allow integration with various existing bio-event annotation schemes, whilst being detailed enough to capture important subtleties in the nature of the meta-knowledge expressed about the event.

4. META-KNOWLEDGE ANNOTATION OF BIO-EVENTS

The annotation scheme presented here is a slightly modified version of our original meta-knowledge annotation scheme [5]. Different types of meta-knowledge are encoded through five distinct dimensions (Figure 2), each of which consists of a set of complete and mutually-exclusive categories, i.e., any given bio-event belongs to exactly one category in each dimension. Our chosen set of annotation dimensions has been motivated by the major information needs of biologists, as discussed earlier. The advantage of using multiple dimensions is that the interplay between the assigned values in each dimension can reveal both subtle and substantial differences in the types of meta-knowledge expressed (see section 4.6).

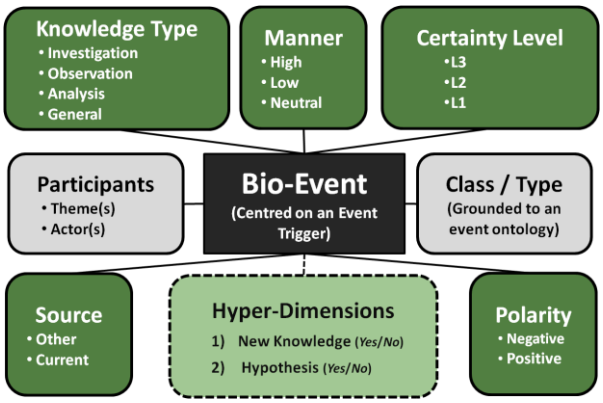

Figure 2. Bio-Event Annotation

Meta-knowledge can be expressed in text in a number of different ways. In the majority of cases, this is through the presence of particular "clue" words or phrases, although other features can also come into play, such as the tense of the verb on which the event is centred, or the relative position of the event within the text.

The annotation task consists of assigning an appropriate value for each dimension, as well as marking the textual evidence for this assignment. In order to minimise the annotation burden, the number of possible categories within each dimension has been kept as small as possible, whilst still respecting important distinctions in meta-knowledge that have been observed during our corpus study. The five meta-knowledge dimensions and their values are described in more detail below.

4.1 Knowledge Type (KT)

This dimension captures the general information content of the event. Each event is classified into one of the following four categories:

Investigation: Enquiries or investigations, which have either already been conducted or are planned for the future, typically marked by lexical clues like *examined*, *investigated* and *studied*, etc.

Observation: Direct observations, often represented by lexical clues like *found* and *observed*, etc. Simple past tense sentences typically also describe observations.

Analysis: Inferences, interpretations, speculations or other types of cognitive analysis, typically expressed by lexical clues like *suggest*, *indicate*, *therefore* and *conclude* etc.

General: Scientific facts, processes, states or methodology. This is the default category for the Knowledge Type dimension.

4.2 Certainty Level (CL)

In scientific text, this dimension is normally only applicable to events whose *KT* corresponds either to *Analysis* or *General*. In the case of *Analysis* events, *CL* encodes confidence in the truth of the event, whilst for *General* events, there is a temporal aspect, to account for cases where a particular process is explicitly stated to

occur most (but not all) of the time, using a marker such as *normally*, or only occasionally, using a marker like *sometimes.* We distinguish three levels of certainty:

L3: No expression of uncertainty or speculation (default category)

L2: High confidence or slight speculation (*Analysis*), event occurs most (but not all) of the time (*General*). Typical lexical markers include *likely* and *probably.* Certain *Analysis* markers also invoke this certainty level, such as *suggest* and *indicate*

L1: Low confidence or considerable speculation (*Analysis*), event occurs infrequently (*General)*; typical lexical markers include *may*, *might* and *perhaps.*

4.3 Source

The source of experimental evidence provides important information for biologists. It can also help in distinguishing new experimental knowledge from previously reported knowledge. Our scheme distinguishes two categories, namely:

Other: The event is attributed to a previous study. In this case, explicit clues (citations or phrases like *previous studies* etc.) are normally present.

Current: The event makes an assertion that can be (explicitly or implicitly) attributed to the current study. This is the default category, and is assigned in the absence of explicit lexical or contextual clues.

4.4 Polarity

This dimension identifies negated events. Although certain bio-event corpora are annotated with this information, it is still missing from others. The indication of whether an event is negated is vital, as the interpretation of a negated event instance is completely opposite to the interpretation of a non-negated (positive) instance of the same event.

We define negation as the absence or non-existence of an entity or a process. Negation is typically expressed by the adverbial *not* and the nominal *no*. However, other lexical devices like negative affixals (*un-* and *in-*, etc.), restrictive verbs (*fail*, *lack*, and *unable*, etc.), restrictive nouns (*exception*, etc.), certain adjectives (*independent*, etc.), and certain adverbs (*without*, etc.) can also be used.

4.5 Manner

This dimension corresponds to indications of the rate, level, strength or intensity of the event described. This can be significant in the correct interpretation of an event. Our scheme distinguishes 3 categories of *Manner*, namely:

High: Typically expressed by adverbs and adjectives like *strongly*, *rapidly* and *high*, etc.

Low: Typically expressed by adverbs and adjectives like *weakly*, *slightly* and *slow*, etc.

Neutral: Default category assigned to all events without an explicit indication of manner.

4.6 Hyper-dimensions

A defining feature of our annotation scheme is that additional information can be inferred by considering combinations of some of the explicitly annotated dimensions. We refer to this additional information as hyper-dimensions of our scheme. At present, we have identified two such hyper-dimensions, as described below.

4.6.1 New Knowledge

A combination of the values of *Source*, *KT* and *CL* dimensions can be used to isolate those events representing new knowledge. For example, events with the *KT* value of *Observation* may correspond to new knowledge, but only if they represent observations from the current study (i.e., *Source=Current*), rather than observations cited from elsewhere. In a similar way, an *Analysis* drawn from experimental results in the current study could be treated as new knowledge, but generally only if it represents a straightforward interpretation of results (i.e. *CL=L3*), rather than something more speculative.

4.6.2 Hypothesis

Events that represent hypotheses can be isolated by considering their values of *KT* and *CL*. Events with a *KT* value of *Investigation* can always be assumed to be a hypothesis. However, if the *KT* value is *Analysis*, then only those events with a *CL* value of L1 or L2 (speculative inferences made on the basis of results) should be considered as hypothesis, to be matched with more definite experimental evidence when available. A value of L3 in this instance would normally be classed as new knowledge, as explained in the previous section.

5. EVALUATION

An initial evaluation of the annotation scheme has been performed through the annotation of 70 abstracts randomly chosen from the GENIA Pathway Corpus, containing a total of 2,603 annotated bio-events. Two of the authors independently annotated these bio-events with meta-knowledge using a comprehensive set of annotation guidelines developed following a detailed analysis of the various bio-event corpora and the output of an initial case study [5]. The remainder of this section discusses the results of this evaluation experiment in more detail.

5.1 Inter-Annotator Agreement

The quality of annotation was assessed using Cohen's kappa [2] to calculate inter-annotator agreement. Table 1 shows the agreement figures for each annotation dimension. The highest value of agreement was achieved for the *Source* dimension, whilst the *KT* dimension yielded the lowest agreement value. Nevertheless, the kappa scores for all annotation dimensions were in the *good* region [4].

5.2 Category Distribution

Knowledge Type: The most prevalent category found in this dimension was *Observation* (45% of events). Only a small fraction of these (4%) was represented by an explicit lexical clue (mostly sensory verbs). In most cases the tense, local context (position within the sentence) or global context (position within the document) were found to be important factors. The second most common category (37% of events) was *General*, of which the majority (64%) were processes or states embedded in noun phrases (such as *c-fos expression*). More than a fifth of the *General* events (22%) expressed known scientific facts, whilst 14% expressed experimental/scientific methods (such as *stimulation* and

Dimension	Cohen's Kappa
KT	0.9017
CL	0.9329
Polarity	0.9059
Manner	0.8944
Source	0.9520

Table 1. Inter-Annotator Agreement

incubation etc.). Explicit lexical clues were found only for facts, but even then in only 1% of cases. *Analysis* was the third most common category of annotated events (16%). Of these events, 44% were deductions (*CL=L3*), whilst the remaining 56% were hedged interpretations (*CL=L1/L2*). All *Analysis* events were marked with explicit lexical clues. The least common category was *Investigation*, comprising 1.5% of all events, all of which were marked with explicit lexical clues.

Certainty Level: *L3* was found to be the most prevalent category, corresponding to 93% of all events. The categories *L2* and *L1* occurred with frequencies of 4.3% and 2.5%, respectively. The relative scarcity of speculative sentences in scientific literature is a well documented phenomenon. Vincze et al. [8] found that less than 18% of sentences occurring in biomedical abstracts are speculative. Similarly, we found that around 20% of corpus events belong to speculative sentences. Since speculative sentences contain non-speculative events as well, the frequency of speculative events is expected to be much less than the frequency of speculative sentences. In accordance with this hypothesis, we found that only 7% of corpus events were expressed with some degree of speculation. We also found that almost all speculated events had explicit lexical clues.

Polarity: Our event-centric view of negation showed just above 3% of the events to be negated. Similarly to speculation, the expected frequency of negated events is lower than the frequency of negated sentences. Another reason for finding fewer negated events is the fact that, in contrast to previous schemes, we draw a distinction between events that are negated and events expressed with *Low* manner. For example, certain words like *limited* and *barely* are often considered as negation clues. However, we consider them as clues for *Low* manner. In all cases, negation was expressed through explicit lexical clues.

Manner: Whilst only a small fraction (4%) of events contains an indication of *Manner*, we found that where present, manner conveys vital information about the event. Our results also revealed that indications of *High* manner are three times more frequent than the indications of *Low* manner. We also noted that both *High* and *Low* manners were always indicated through the use of explicit clues.

Source: Most (99%) of the events were found to be of the *Current* category. This is to be expected, as authors tend to focus on current work in within abstracts. It is envisaged, however, that this dimension will be more useful for analyzing full papers.

Hyper-dimensions: Almost 57% of the events represent *New Knowledge*, and just above 8% represent *Hypotheses*.

6. CONCLUSION AND FUTURE WORK

The recent advent of event-centred text mining approaches mandates the need for correct and consistent interpretation of textual events. We have presented a new approach to address this problem in the domain of biomedical research literature. The cornerstone of our approach is a meta-knowledge annotation scheme that captures the key information required for the correct interpretation of bio-events [5]. An initial evaluation experiment has illustrated high inter-annotator agreement and a sufficient number of annotations along each category in every dimension. The highly favourable results of this experiment have confirmed the feasibility and soundness of the annotation scheme, and have paved the way for a large scale annotation effort involving multiple independent (i.e. non-author) annotators.

We are currently in the process of creating a large corpus of meta-knowledge enriched bio-events. This corpus will consist of three sub-corpora, which have previously been annotated with different types of bio-events, namely GENIA, GREC and a small corpus of full papers.

7. ACKNOWLEDGMENTS

The work described in this paper has been funded by the Biotechnology and Biological Sciences Research Council through grant numbers BBS/B/13640, BB/F006039/1 (ONDEX).

8. REFERENCES

[1] Ashburner, M., Ball, C. A., Blake, J. A. , Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. 2000. Gene ontology: tool for the unification of biology. *Nature Genetics* 25:25-29.

[2] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 20, 37–46.

[3] Kim, J., T. Ohta and Tsujii, J. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics* 9:10

[4] Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills.

[5] Nawaz, R., Thompson, P., McNaught, J. and Ananiadou, S. (2010). Meta-Knowledge Annotation of Bio-Events. In *Proceedings of LREC 2010*, 2498-2507.

[6] Oda, K., Kim, J., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y. and Tsujii, J. 2008. New challenges for text mining: mapping between text and manually curated pathways. BMC Bioinformatics 9(Suppl 3): S5.

[7] Thompson, P., Iqbal, S., McNaught, J. and Ananiadou, S. 2009. Construction of an annotated corpus to support biomedical information extraction. *BMC Bioinformatics* 10: 349

[8] Vincze, V., Szarvas, G., Farkas, R., Mora, G. and Csirik, J. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 9(Suppl 11): S9.

[9] Waard, A. de., Shum, B., Carusi, A., Park, J., Samwald M. and Sándor, Á. 2009. Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims. In *Proceedings of the Workshop on Semantic Web Applications in Scientific Discourse.* Available at: http://oro.open.ac.uk/18563/

[10] Wilbur, W.J., Rzhetsky, A. and Shatkay, H. 2006. New directions in biomedical text annotations: definitions, guidelines and corpus construction. BMC Bioinformatics 7: 356.