Meta-Knowledge Annotation of Bio-Events

Raheel Nawaz[1], Paul Thompson[1,2], John McNaught[1,2], Sophia Ananiadou[1,2]
[1]School of Computer Science, University of Manchester, UK
[2]National Centre for Text Mining, University of Manchester, UK
E-mail: nawazr@cs.man.ac.uk, paul.thompson@manchester.ac.uk, john.mcnaught@manchester.ac.uk,
sophia.ananiadou@manchester.ac.uk

Abstract

Biomedical corpora annotated with event-level information provide an important resource for the training of domain-specific information extraction (IE) systems. These corpora concentrate primarily on creating classified, structured representations of important facts and findings contained within the text. However, bio-event annotations often do not take into account additional information (*meta-knowledge*) that is expressed within the textual context of the bio-event, e.g., the pragmatic/rhetorical intent and the level of certainty ascribed to a particular bio-event by the authors. Such additional information is indispensable for correct interpretation of bio-events. Therefore, an IE system that simply presents a list of "bare" bio-events, without information concerning their interpretation, is of little practical use. We have addressed this sparseness of meta-knowledge available in existing bio-event corpora by developing a multi-dimensional annotation scheme tailored to bio-events. The scheme is intended to be general enough to allow integration with different types of bio-event annotation, whilst being detailed enough to capture important subtleties in the nature of the meta-knowledge expressed about different bio-events. To our knowledge, our scheme is unique within the field with regards to the diversity of meta-knowledge aspects annotated for each event.

1. Introduction

Biomedical corpora annotated with event-level information, (e.g., Pyysalo et al., 2007; Kim et al., 2008; Thompson et al., 2009), provide an important resource for the training of domain-specific information extraction (IE) systems. These corpora concentrate primarily on creating classified, structured representations of important facts and findings contained within the text. As an example, consider the simple sentence shown in Figure 1.

The results suggest that the narL gene product activates the nitrate reductase operon.

Figure 1: A Simple Sentence from a Biomedical Abstract

This sentence contains a single *bio-event*, described by the verb *activates*. Figure 2 shows a typical structured representation of this bio-event.

EVENT-TRIGGER: *activates*
EVENT-TYPE: positive_regulation
THEME: *nitrate reductase operon:* operon
CAUSE: *narL gene product*: protein

Figure 2: Typical Structured Representation of the Bio-Event mentioned in Figure 1

Entities involved in the bio-event (i.e., the subject and object of *activates*) have been assigned both biological *named entity* types (i.e., protein and operon) as well as *semantic roles* (i.e., cause and theme) indicating their contribution towards the meaning of the event.

A system trained to extract such representations, and which allows searches to be performed over these bio-events, can help biologists to locate relevant information much more quickly than is possible using the traditional method of keyword searches over unstructured documents. For example, *semantic queries* can be formulated that specify various types of semantic restrictions on the bio-events to be retrieved. The parameters for these semantic queries can include event types, semantic role labels and named entity types etc. (Miyao et al., 2006).

What is often missed by bio-event annotation (and hence in systems trained on the annotated corpora) is the additional information expressed within the textual context of the bio-event. However, correct interpretation of bio-events is not possible without such additional information. For example, the first part of the example sentence in Figure 1 (i.e., "*The results suggest that"*) indicates that the occurrence of the positive regulation event is not a definite fact; instead, it is based on an analysis of experimental results. Therefore, an IE system that simply presents a list of "bare" bio-events, without information concerning their interpretation, is of little practical use.

2. Meta-Knowledge

Typical tasks in which biologists have to search and review the literature include building and updating models of biological processes, such as pathways (Oda et al., 2008), and curation of biological databases (Ashburner et al., 2000; Yeh et al., 2003). Central to both of these tasks is the identification of *new knowledge* that can enhance these resources, e.g. to build upon an existing, but incomplete model of a biological process

(Lisacek et al., 2005) or to ensure that the database is kept up to date. Any new knowledge added should be supported though evidence, which could include linking hypotheses with experimental findings. It is also important to take into account inconsistencies or contradictions reported in the literature.

Thus, in addition to recognising bio-events themselves, an IE system that is to be fully useful to biology experts in scenarios such as the above should have the ability to recognise and present *meta-knowledge* about the bio-events to its users. Such knowledge includes identifying the author's *rhetorical/pragmatic intent* behind the bio-event (de Waard et al., 2009), e.g., whether the bio-event represents a hypothesis, accepted knowledge or new experimental knowledge. The nature of the new knowledge may also be important, i.e., it could correspond to directly observed evidence, or it may represent an inference drawn from experimental results. In the latter case, the author's level of *certainty* towards the bio-event may provide important information regarding the perceived reliability of the inference.

To this end, we have defined a new annotation scheme for enriching bio-events with these and other types of meta-knowledge, with the aim of facilitating the training of more useful systems in the context of various IE and textual inference (TI) tasks. Whilst the scheme has been designed to capture and classify a wide range of useful information, its suitability for application to existing bio-event corpora has also been a major design consideration. To our knowledge, our scheme is unique within the field with regards to the diversity of meta-knowledge aspects annotated for each event.

3. Lexical Markers of Meta-Knowledge

Several previous studies have looked at how information in biomedical texts can be classified to aid in its interpretation. One thread of research has studied the lexical markers (i.e., words or phrases) which can accompany statements to indicate their intended interpretation. Rizomilioti (2006) examined biology research articles (amongst others) in order to compile lists of lexical items which indicate different levels of certainty. The analysis carried out by Hyland (1996) on cell and molecular biology articles provided a more detailed analysis of lexical items used in *hedges* (i.e., speculative statements), including those that denote deductions and sensory (i.e., visual) evidence, in addition to speculations. Our own previous work (Thompson et al., 2008) also concerned a comparable categorization of lexical markers, although, in contrast to other studies, we took a multi-dimensional approach to the categorization, acknowledging that several types of important information may be expressed through different words in the same sentence. As an example, let us consider the example sentence in Figure 3. The author's pragmatic/rhetorical intent towards the statement that *the catalytic role of these side chains is associated with their interaction with the DNA substrate* is encoded by the word *indicate*, which shows that the statement represents an analysis of the evidence stated at the beginning of the sentence, i.e., that the mutations at positions 849 and 668 have DNA-binding properties. Furthermore, the author's *certainty level* (i.e., their degree of confidence) towards this analysis is shown by the word *may*. Here, the author is uncertain about the validity of their analysis.

> *The DNA-binding properties of mutations at positions 849 and 668 may indicate that the catalytic role of these side chains is associated with their interaction with the DNA substrate.*

Figure 3: Example Sentence

Whilst our previous work served to demonstrate that the different aspects of meta-knowledge that can be specified within texts require a multi-dimensional analysis to correctly capture their subtleties, we also concede that taking a purely lexical approach to the recognising meta-knowledge in texts (i.e., simply looking for words from these lists that co-occur in the same sentences as events of interest) is not sufficient. The reasons for this include:

a) The presence of a particular marker does not guarantee that the "expected" interpretation can be assumed (Sándor, 2007). Some markers may have senses which vary according to their context. As noted by Hyland (2005, p.125), "Every instance should ... be studied in its sentential co-text".

b) Not all types of meta-information are indicated through explicit markers. Mizuta & Collier (2004), who annotated *rhetorical zones* in texts based on the scheme proposed by Teufel et al. (1999), found that different types of zones in texts may be indicated not only through explicit lexical markers but also through features such as the main verb in the clause and the position of the sentence within the article or abstract.

It is thus important to perform annotation on *all* relevant instances, regardless of the presence of lexical markers. This will allow systems to be trained that can learn to determine the correct meta-knowledge category, even when lexical markers are not present. In addition, the trained system should be able to discriminate the contexts in which the presence of particular lexical markers can reliably predict meta-knowledge categories.

4. Existing Corpora with Meta-Knowledge Annotations

Various corpora of biomedical literature (abstracts and/or full papers) have been produced that feature some degree of meta-knowledge annotation. These corpora vary in both the richness of the annotation added, and the type / size of the units at which the meta-knowledge annotation has been performed. Taking the unit of annotation into account, we can distinguish between annotations that apply to continuous text-spans, and annotations that have been performed at the event level.

4.1 Text-Span Annotation

Existing meta-knowledge annotations performed on continuous text-spans generally only cover a single dimension of annotation, corresponding to either speculation/certainty level, (e.g., Light et al., 2004; Medlock & Briscoe, 2007; Vincze et al., 2008) or general information content/rhetorical intent, e.g., *background, methods, results, insights*. This latter type of annotation has been attempted both on abstracts , (e.g., McKnight & Srinivasan, 2003; Ruch et al., 2007) and full papers, (e.g. Teufel et al., 1999; Langer et al., 2004; Mizuta & Collier, 2004), with the number of distinct annotation categories varying between 4 and 14. Liakata et al. (2010) also perform this type of annotation, but they create further specialisations by annotating properties of each concept type (e.g. new or previous work).

An issue with using sentences or *zones* (Teufel et al., 1999) as the unit of annotation is that a single sentence may express several types of information or rhetorical functions, e.g., both an experimental method and the results of applying this method. Equally, an expression of speculation may apply only to part of a sentence. This issue is addressed by Wilbur et al. (2006), whose scheme sits somewhere between sentence and event level annotation, in that it applies to sentence *fragments*, which are created on the basis of changes in the meta-knowledge expressed. The scheme consists of multiple annotation dimensions which capture aspects of both certainty and rhetorical/pragmatic intent, amongst other things. Later work on this scheme provides evidence that training a system to automatically annotate these dimensions is highly feasible (Shatkay et al., 2008).

4.2 Event Level Annotation

Annotation of detailed meta-knowledge at the event level has so far attracted less attention than text-span annotation. However, as the recognition of bio-events is central to many biomedical text mining applications, annotation of meta-knowledge at this level is more practically useful, allowing systems to be trained to assign more specific and semantically precise information to particular bio-events than is possible using sentence-level annotation.

To our knowledge, only the GENIA corpus (Kim et al., 2008) contains certainty-level annotation at the bio-event level, whilst annotation of rhetorical/pragmatic intent at this level has not been attempted in any corpora. Although Sanchez-Graillet & Poesio (2007) propose a multi-dimensional annotation scheme for protein-protein interaction events, including certainty, manner and direction, their reported corpus only contains one dimension of annotation, i.e., polarity. Polarity has also been annotated in other event corpora within the domain, (e.g., Pyysalo et al., 2007; Kim et al., 2008).

We have addressed the current sparseness of meta-knowledge available in existing bio-event corpora by developing a multi-dimensional annotation scheme tailored to bio-events. The scheme is intended to be general enough to allow integration with different bio-event annotation schemes, whilst being detailed enough to capture important subtleties in the nature of the meta-knowledge expressed about the event.

As our previous work (Thompson et al., 2008) showed that lexical meta-knowledge markers can be a key factor in determining the values of certain meta-knowledge dimensions, our current annotation scheme proposes to annotate such markers, when they are present. This is in contrast to most of the other annotation schemes discussed above.

5. Annotation Scheme

Our scheme consists of six meta-knowledge dimensions, each with 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 discussed earlier, i.e., the ability to distinguish between different intended interpretations of events, including hypotheses, existing knowledge or new experimental knowledge, and, in the latter case, the nature of the experimental knowledge.

The scheme combines various aspects of existing annotation schemes, with appropriate modifications that take into account our detailed study of a large number of event-annotated abstracts within our specific domain. In addition, 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 advantage of using a multi-dimensional scheme is that the interplay between different values of each dimension can reveal both subtle and substantial differences in the types of meta-knowledge expressed in the surrounding text. Therefore, in most cases, the exact rhetorical/pragmatic intent of an event can only be determined by considering a combination of the values of different dimensions.

Figure 4 provides an overview of the annotation scheme. The boxes with the light-coloured (grey) background correspond to information that is common to most bio-event annotation schemes, i.e., the participants in the event, together with an indication of the class or type of the event. The boxes with the darker (green) backgrounds correspond to our proposed meta-knowledge annotation dimensions and their possible values. The remainder of this section provides brief details of each annotation dimension.

5.1 Knowledge Type (KT)

This dimension is responsible for capturing the general information content of the event. Whilst less detailed than some of the previously-proposed sentence-level schemes, its purpose is to form the basis of distinguishing between the most critical types of rhetorical/pragmatic intent, according to the needs of biologists. Each event is thus classified into one of the following 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. Such events can normally be interpreted as hypotheses.
- Observation: Direct observations, often represented by lexical clues like *found*, *observed* and *report*, etc. Simple past tense sentences typically also describe observations. Such events represent experimental knowledge.
- Analysis: Inferences, interpretations, speculations or other types of cognitive analysis, typically expressed by lexical clues like *suggest*, *indicate*, *therefore* and *conclude* etc. Such events, if they are interpretations or reliable inferences based on experimental results, can also constitute another type of (indirect) experimental knowledge. Weaker inferences or speculations, however, may be considered as hypotheses which need further proof through experiments.
- General: Scientific facts, processes or methodology (default category). Many of these events will correspond to generally accepted knowledge.

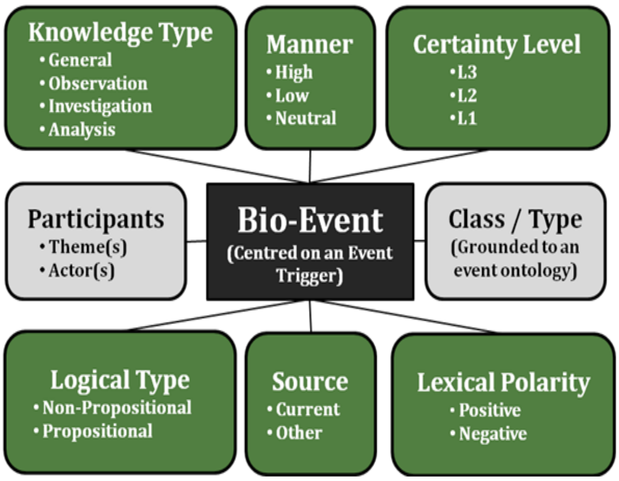

Figure 4: Bio-Event Annotation

The above category descriptions show that, taken alone, *Knowledge Type* is not always sufficient to determine the complete rhetorical/pragmatic intent behind the event. For example, an *Observation* represents experimental knowledge or evidence. However, for certain tasks, it is important to be able to distinguish *new* evidence reported in a paper from that which has been reported elsewhere. This is taken care of by the *Source* attribute, which is described in section 5.3. In the case of *Analysis* events, the perceived reliability of the event is important in determining whether the event can be treated as experimental knowledge (in which case the *Source* attribute may also come into play), or as a hypothesis. Within our scheme, event reliability or confidence is encoded by the *Certainty Level* attribute, described in section 5.2.

5.2 Certainty Level (CL)

The value of this dimension is almost always indicated through the presence of an explicit lexical marker. In scientific literature, it 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.* Events corresponding to direct *Observations* are not open to judgements of certainty, nor are *Investigation* events, which refer to things which have not yet happened or have not been verified.

Regarding the choice of values for the *CL* dimension, there is an ongoing discussion as to whether it is indeed possible to partition the epistemic scale into discrete categories (Rubin, 2007). However, the use of a number of distinct categories is undoubtedly easier for annotation purposes and has been proposed in a number of previous schemes. Although recent work has suggested the use of four or more categories (Shatkay et al., 2008; Thompson et al., 2008), our initial analysis of bio-event corpora has shown that only three levels of certainty seem readily distinguishable for bio-events. This is in line with Hoye (1997), whose analysis of general English showed that there are at least three articulated points on the epistemic scale.

We have chosen to use numerical values for this dimension, in order to reduce potential annotator confusions or biases that may be introduced through the use of labels corresponding to particular lexical markers of each category, such as *probable* or *possible*, and also to account for the fact that slightly different interpretations apply to the different levels, according to whether the event has a *KT* value of *Analysis* or *General*.

- L3: No expression of uncertainty or speculation (default category)
- L2: High confidence or slight speculation.
- L1: Low confidence or considerable speculation; typical lexical markers include *may*, *might* and *perhaps.*

Events corresponding to the *KT* category of *Analysis* are highly likely to correspond to new experimental knowledge when the certainty level is either *L3* (generally, interpretations of results) or *L2* – (normally, high confidence inferences made on the basis of results). Analyses with a certainty level of *L1* would normally be too tentative to be classed as new experimental knowledge, and rather should be treated as hypotheses to be matched with more definite experimental evidence when available.

5.3 Source

The source of experimental evidence provides important information for biologists. This is demonstrated by its annotation during the creation of the Gene Ontology (Ashburner et al., 2000) and in the corpus created by

Wilbur et al. (2006). As explained above, the *Source* dimension 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 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.

5.4 Lexical Polarity (LP)

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.

5.5 Manner

Events may be accompanied by a word or phrase which provides an indication of the rate, level, strength or intensity of the interaction. We refer to this as the *Manner* of the event. Information regarding manner has not been annotated in the majority of existing bio-event corpora, but yet the presence of such words can be significant in the correct interpretation of the 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.

5.6 Logical Type (LT)

This dimension aims to determine whether the event represents a proposition or not. We define an event as propositional if the text provides explicit information about its "truth value". Propositional events typically represent the main assertion(s) in a given clause, whilst non-propositional events are typically those which correspond to arguments of propositional events, and are normally centred on nominalised verbs.

The surface representations of non-propositional events typically do not provide enough information to allow annotation along the *CL*, *Manner*, *Polarity* and *Source* dimensions.

6. Annotation Example

This section aims to further clarify the annotation scheme through a set of example sentences (Figure 6) and their associated annotations (Table 1). Two bio-events occur in these sentences. Event E1 represents the expression of an arbitrary gene *X*, whilst event E2 represents the positive regulation of E1 by an arbitrary protein *Y*. Figure 5 shows the typical structured representation of these events:

EVENT-ID:	E1
EVENT-TYPE:	gene_expression
THEME:	*X :* gene

EVENT-ID:	E2
EVENT-TYPE:	positive_regulation
THEME:	*E1:* event
CAUSE:	*Y*: protein

Figure 5: Structured Representation of E1 and E2

Although each example sentence contains an instance of one or both of the same bio-events (E1 and E2), their interpretations vary according to the sentential context. More importantly, without the annotation of meta-knowledge information, the events extracted from each sentence would be identical, and the differences in meaning expressed within the sentential context would be lost.

S1 = We found that Y activates the expression of X
S2 = We examined the effect of Y on expression of X
S3 = These results suggest that Y has no effect on expression of X
S4 = Y is known to increase expression of X
S5 = Addition of Y slightly increased the expression of X
S6 = These results suggest that Y might affect the expression of X
S7 = Significant expression of X was observed
S8 = Previous studies have shown that Y activates the expression of X

🟨 KT Clue 🟫 Manner Clue 🟩 CL Clue
🟥 Source Clue 🟦 Polarity Clue

Figure 6: Example Sentences

In sentence S1, the presence of the word *found* indicates that the positive regulation event (E2) being reported is an experimental outcome, i.e., an observation. Therefore, the *KT* value for the event should be *Observation*. The Source value is set to the default of *Current*, as there is no evidence in the surrounding context that the event refers to previous work. As mentioned above, this combination of *KT* and *Source* values indicates that the information contained in the event can be flagged as new knowledge.

Sentence ID	E1						E2					
	Knowledge Type	Certainty Level	Manner	Polarity	Source	Logical Type	Knowledge Type	Certainty Level	Manner	Polarity	Source	Logical Type
S1	General	-	-	-	-	Non-Prop	Observation	L3	Neutral	Positive	Current	Prop
S2	General	-	-	-	-	Non-Prop	Investigation	-	-	-	-	Non-Prop
S3	General	-	-	-	-	Non-Prop	Analysis	L2	Neutral	Negative	Current	Prop
S4	General	-	-	-	-	Non-Prop	General	L3	Neutral	Positive	Current	Prop
S5	General	-	-	-	-	Non-Prop	Observation	L3	Low	Positive	Current	Prop
S6	General	-	-	-	-	Non-Prop	Analysis	L1	Neutral	Positive	Current	Prop
S7	Observation	L3	High	Positive	Current	Prop	-	-	-	-	-	-
S8	General	-	-	-	-	Non-Prop	Observation	L3	Neutral	Positive	Other	Prop

Table 1: Meta-Knowledge Annotation of Sentences S1-S8

The interpretation of E2 in S8 is very similar to S1; the presence of the word *shown* explicitly indicates that E2 is an observation (i.e., *KT* = *Observation*). However, the use of *previous studies* at the start of the sentence indicates that these results were originally reported outside the current paper (i.e., Source value = *Other*). Therefore, whilst this information constitutes experimental evidence, it does not correspond to *new* knowledge. Sentence S4 also contains an instance of E2 with a similar interpretation to S1 and S8. However, the word "*known*" indicates that E2 is a well established fact within the field (i.e., *KT* = *General*).

Whilst there are subtle differences in the interpretation of E2 in S1, S4 and S8, they all have in common that the event is expressed as a definite fact. In this respect, the instance of E2 in S2 is quite different. Here, the presence of the word *examined* indicates that the positive regulation event is under examination (i.e., *KT* = *Investigation*), and so its truth value is unknown. Thus, it would be incorrect for a text mining system to present E2, in this context, as a definite fact. Rather, it could be considered a hypothesis.

In sentence S6, there is yet a different interpretation of E2. The use of the verb *suggest* indicates that an instance of E2 has been reported based on an inference drawn from results (i.e., *KT* = *Analysis*). However, the presence of the word *might* indicates that this inference is a speculation (i.e., *CL* = *L1*), and hence is too weak to be treated as new experimental evidence.

Sentence S3 is similar to S6, in that it also uses the word *suggests* to indicate an inference (i.e., *KT* = *Analysis*). The lack of an accompanying *CL* marker shows that the author is fairly confident about this inference, and so it can be considered reliable enough to be treated as new knowledge. However, the conclusion is different from S6: the authors use the word *no* to indicate that E2 *does not* occur (i.e., *Polarity* = *Negative*).

In sentence S5, the word *slightly* provides explicit information about the intensity of E2 (i.e., *Manner* = *Low*). The recognition of such information about events may be important, for example, when performing a comparison of the results of different experimental methods. In sentence S7, the intensity of the expression event is also indicated, this time by the word *significant* (i.e., *Manner* = *High*).

7. Case Study

We have conducted a small annotation case-study on 715 randomly chosen bio-events from the GENIA event corpus to verify the suitability of our annotation scheme for application to the existing bio-event corpora. Table 2 shows the distribution of events among the categories of each annotation dimension. A summary of our findings is as follows:

<u>Knowledge Type:</u> *General* (58%), was the most prevalent category, although it was rarely marked by lexical clues. Most events in this category (92%) corresponded to processes embedded in non-propositional text fragments (such as *c-fos expression*), and a small fraction (8%) were known scientific facts. Almost a third of events belonged to the *Observation* category. Of these, 24% were represented by an explicit lexical clue. In the other cases, either tense or position within the abstract were found to be important features. Events in the *Analysis* and *Investigation* categories were all marked with lexical clues.

<u>Certainty Level:</u> 92% of events belonged to category *L3*, 5% to *L2* and 3% to *L1*. The relative paucity of speculative sentences in biomedical literature is a well documented phenomenon (Thompson et al., 2008; Vincze et al., 2008). We found that *events* expressed

with some degree of speculation (*L2* + *L1*) are even more scarce, because speculative sentences often contain non-speculative events as well. We also noted that certain words (like *suggest*, *speculate* etc.) can be used as clues for values of both CL and KT categories.

Dimension	Category	No of Events	% of Events
Knowledge Type (KT)	Analysis	71	10%
	Investigation	22	3%
	Observation	210	29%
	General	412	58%
Certainty Level (CL)	L1	25	3%
	L2	33	5%
	L3	657	92%
Polarity	Negative	47	7%
	Positive	668	93%
Manner	High	20	3%
	Low	8	1%
	Neutral	687	96%
Source	Current	703	98%
	Other	12	2%
Logical Type (LT)	Propositional	304	43%
	Non-Propositional	411	57%

Table 2: Annotation Results

Polarity: Vincze et al. (2008) found that less than 14% sentences occurring in biomedical abstracts are negative. However, our event-centred view of negation showed that more than 19% of events belong to sentences containing some kind of negation, although only 7% of events were found to be negated.

Manner: Whilst only a small fraction (4%) of events contains an indication of manner, we noted that, where present, the manner conveys vital information about the event. Our results also revealed that indications of *High* manner are much more frequent than the indications of *Low* manner.

Source: Most (98%) of the events were found to be of the *Current* category. This is partly because authors tend not to use citations in abstracts. It is envisaged, however, that this dimension will be more useful for analyzing full papers.

Logical Type: 43% of annotated events were *Propositional* and the remaining 57% were *Non-Propositional*. This high number of non-propositional events is unsurprising: we note that non-propositional events tend to be centred on nominalised verbs, and in Thompson et al. (2009) it was shown that around half of all events in the gene-regulation domain are centred on nominalised verbs. Our results show that all non-propositional events belong to the *KT* categories of *General* or *Investigation*.

8. Conclusion and Future Work

We have presented a multi-dimensional meta-knowledge annotation scheme for bio-events. The scheme captures key information regarding the correct interpretation of bio-events, which is not currently annotated in existing bio-event corpora, but which we have shown to be critical in a number of text mining tasks undertaken by biologists. The results of our case-study have confirmed the feasibility of the annotation scheme for application to existing bio-event corpora, with the proposed categories in all dimensions having been annotated, at least to some extent.

We are currently in the process of producing as larger, manually annotated corpus in which documents from existing bio-event corpora are being annotated according to our meta-knowledge annotation scheme. Once completed, this corpus will serve as a useful resource for the development of automatic meta-knowledge annotation systems.

9. Acknowledgements

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)

10. References

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, pp 25--29.

de Waard, A., 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/

Hoye, L. (1997). *Adverbs and Modality in English*. London & New York: Longman

Hyland, K. (1996). Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication* 13(2), pp. 251--281.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. London: Continuum

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

Langer, H., Lungen, H. and Bayerl, P. S. (2004). Text type structure and logical document structure. In *Proceedings of the ACL Workshop on Discourse Annotation*, pp. 49 --56

Liakata, M.,Teufel, S., Siddharthan, A. and Batchelor., C. (2010) Corpora for conceptualisation and zoning of scientific papers. To appear in *Proceedings of LREC 2010*.

Light, M., Qui, X. T. and Srinivasan, P. (2004). The language of bioscience: Facts, speculations, and statements in between. In *Proceedings of the BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, pp 17--24.

Lisacek, F., Chichester, C., Kaplan, A. and Sandor, A. (2005). Discovering Paradigm Shift Patterns in Biomedical Abstracts: Application to Neurodegenerative Diseases. In *Proceedings of SMBM 2005*, pp 212--217

McKnight, L. and Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *Proceedings of the 2003 Annual Symposium of AMIA*, pp 440--444.

Medlock, B. and Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. In *Proceedings of ACL 2007*, pp. 992-- 999.

Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. and Tsujii, J. (2006). Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. In *Proceedings of COLING-ACL 2006*, pp 1017--1024.

Mizuta, Y. and Collier, N. (2004). Zone identification in biology articles as a basis for information extraction. In *Proceedings of the joint NLPBA/BioNLP Workshop on Natural Language for Biomedical Applications*, pp. 119--125.

Oda, K., Kim, J.-D., 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.

Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J. and Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. *BMC Bioinformatics* 8: 50.

Rizomilioti, V. (2006). "Exploring Epistemic Modality in Academic Discourse Using Corpora." *Information Technology in Languages for Specific Purposes* 7, pp 53--71

Rubin, V. L. (2007). Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. In *Proceedings of NAACL-HLT 2007, Companion Volume***,** pp. 141--144.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., Gobeill, J., Pillet, V., Rebholz-Schuhmann, D. and Lovis, C. (2007). Using argumentation to extract key sentences from biomedical abstracts. *International Journal of Medical Informatics* 76(2-3), pp. 195--200.

Sanchez-Graillet, O. and Poesio M. (2007). Negation of protein-protein interactions: analysis and extraction. *Bioinformatics* 23(13), pp. i424--i432

Sándor, Á. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée* 200(2), pp 97--109.

Shatkay, H., Pan, F., Rzhetsky, A. and Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics* 24(18), pp. 2086--2093.

Teufel, S., Carletta, J. and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of EACL 1999*, pp. 110--117.

Teufel, S, Siddharthan, A. and Batchelor, C (2009) Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP-09*, pp. 1493--1502

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

Thompson, P., Venturi, G., McNaught, J., Montemagni, S. and Ananiadou, S. (2008). Categorising Modality in Biomedical Texts. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pp 27—34.

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.

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

Yeh, A. S., Hirschman, L. and Morgan, A. A. (2003). Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. Bioinformatics 19(Suppl 1), pp. i331--i339.