Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers

Raheel Nawaz[1], Paul Thompson[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: raheel.nawaz@cs.man.ac.uk, paul.thompson@manchester.ac.uk, sophia.ananiadou@manchester.ac.uk

Abstract

Biomedical literature contains rich information about events of biological relevance. Event corpora, containing classified, structured representations of important facts and findings contained within text, provide an important resource for the training of domain-specific information extraction (IE) systems. Such corpora pay little attention to the interpretation of events, e.g., whether an event describes a fact or an analysis of results, whether there is any speculation surrounding the event, etc. These types of information are collectively referred to as *meta-knowledge*. As previous work, an annotation scheme to enrich event corpora with meta-knowledge was designed to facilitate the training of more sophisticated IE systems, and was applied to the complete GENIA Event corpus of biomedical abstracts. In this paper, we describe a case study in which four full papers annotated with GENIA events have been manually enriched with meta-knowledge annotation. We analyse the annotation results, and compare them with the previously annotated abstracts.

Keywords: meta-knowledge, annotation, events, information extraction, biomedical literature

1. Introduction

Due to the rapid growth in the body of scientific literature, it is becoming increasingly important to move beyond simple keyword-based searching to more sophisticated methods that can help researchers to isolate information of interest from a potential mountain of relevant documents. Accordingly, text mining has been receiving increasing interest within the biomedical field (Zweigenbaum et al., 2007). In particular, information extraction (IE) systems produce structured, template-like representations of important facts and findings within documents, called *events*. The extracted events can form the basis of sophisticated semantic search systems, in which users specify search criteria through the (partial) completion of a structured template, which is matched against the extracted events.

IE systems are sensitive to the features of the text on which they operate, and relevant event types vary between domains. Accordingly, such systems must be adapted to deal with specific domains. The usual method of adaptation is the application of machine-learning methods to annotated corpora, e.g. (Soderland, 1999; Califf & Mooney, 2003). In the biomedical field, several corpora annotated with events have been produced, most notably the GENIA event corpus (Kim et al., 2008), the BioInfer corpus (Pyysalo et al., 2007) and the GREC corpus (Thompson et al., 2009). Research into event extraction systems was greatly boosted by the BioNLP'09 shared task on event extraction, in which 24 teams participated (Kim et al., 2009).

Until recently, most event corpora, and thus the systems trained on them, dealt exclusively with abstracts from small subdomains of molecular biology. However, the development of systems that automatically analyse full papers is also vital, given that less than the 8% of scientific claims occur in abstracts (Blake, 2010). However, since there are significant structural and linguistic differences between full papers and abstracts (Cohen et al., 2010), adapting text mining technology from abstracts to full papers presents significant challenges. In terms of event extraction, an effort to move beyond the previous constraints is described in Pyysalo et al. (2010), which concerned the extraction of events from full papers in a new domain, i.e. infectious diseases. This theme was continued in the BioNLP Shared Task 2011 (Kim et al., 2011a), which included tasks relating to four different domains. The original corpus from the BioNLP'09 shared task (derived from the GENIA event corpus) was extended with a small number of full papers annotated according to the same event scheme, to allow evaluation of event extraction technology on full papers (Kim et al., 2011b).

The focus of the annotation in most event corpora is on locating appropriate events in texts, assigning types to them and identifying event participants. However, detailed information about how the events are to be interpreted according to their textual context is usually missing from the annotations. Such information is termed as "meta-knowledge" (Nawaz et al., 2010). Very basic meta-knowledge information is included in most existing corpora, e.g., negated events are identified in BioInfer corpus, whilst negation and basic speculation information are present in the GENIA corpus and the two related corpora from the two BioNLP shared tasks. Such basic meta-knowledge is, however, not sufficient to distinguish between events that express the following types of meta-knowledge:

- Accepted facts vs. experimental findings.
- Hypotheses vs. interpretations of experimental results.
- Previously reported findings vs. new findings.

Previously, an annotation scheme tailored enriching biomedical event corpora with detailed meta-knowledge along five different dimensions was defined (Nawaz et al., 2010). A slightly modified version of the meta-knowledge scheme was subsequently applied to the GENIA Event corpus (1000 MEDLINE abstracts, containing 36,858 events) (Thompson et al., 2011).

In line with the extension of event extraction systems to deal with full papers, it is important to ensure that meta-knowledge can also be assigned to events in full texts. As a first step, we have performed a case study in which we have applied our meta-knowledge scheme to 4 event-annotated full papers. In this paper, we analyse the outcomes of this new meta-knowledge annotation effort, and compare the results to those obtained for abstracts in the GENIA event corpus. It is our intention that insights gained will help to feed into the design of systems that can automatically assign meta-knowledge at the level of full papers as well as abstracts.

2. Event-Based Text Mining

The process of event annotation normally consists of the identification of an event trigger and event participants, and the assignment of types/categories to each of these. The *event-trigger* is a word or phrase in the sentence that indicates the occurrence of the event (often a verb or nominalisation). The *event-type* (generally assigned from an ontology) categorises the type of information expressed by the event. The event participants, i.e., entities or other events that contribute towards the description of the event, are often categorised using semantic role labels such as *cause* and *theme*. Usually, semantic types (e.g. *gene*, *protein*, etc.) are also assigned to the named entities (NEs) participating in the event.

In order to illustrate this typical event representation, consider the following sentence from GENIA Event corpus (PMID: 3035558):

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

Figure 1 shows the typical structured representation of the biomedical event described in this sentence.

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

Figure 1: Typical representation of a bio-event

The automatic recognition of such events allows users to create structured queries, on which different kinds of restrictions can be specified to restrict the types of events to be retrieved (Miyao et al., 2006). These restrictions may concern the type of event to be retrieved, the types of participants that should be present in the event or the values of these participants, in terms of either specific strings or NE types.

3. Meta-Knowledge Annotation Scheme

Our event-based meta-knowledge scheme aims to capture as much useful information as possible about individual events from their textual context, to support the training of enhanced event-based search systems. Such enhanced systems could improve the efficiency of tasks such as building and updating models of biological processes, e.g., 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, i.e. experimental findings or conclusions that relate to the current study, and which are stated with a high degree of confidence. Meta-knowledge identification is also useful when checking for inconsistencies or contradictions in the literature, since the meta-knowledge values assigned to two otherwise identical events can affect their interpretation in both subtle and significant ways.

The scheme consists of multiple annotation dimensions to capture different aspects of meta-knowledge. For each dimension, a single category is assigned from a fixed set of possible values. If the category of a given dimension is assigned based on the presence of a particular word or phrase in the sentence, this is also annotated as a “clue”. The scheme was inspired by previous multi-dimensional efforts to assign meta-knowledge to continuous text spans, e.g. (Wilbur et al., 2006; Liakata et al., 2010). The feasibility of automating annotation according to both of these schemes has subsequently been demonstrated (Shatkay et al., 2008; Liakata et al., 2012).

In contrast to the two schemes mentioned above, which concern the annotation of continuous text spans, our meta-knowledge annotation scheme (Thompson et al, 2011) is the first that is specifically tailored to the enrichment of event annotations. In addition to allowing several distinct types of information to be encoded about events, the multi-dimensional nature of the scheme allows the interplay between the different dimension values to be used to derive further useful information (*hyper-dimensions*) regarding the interpretation of the event. The scheme is summarized in Figure 2. A brief overview of the dimensions of our scheme and their possible values are provided below. Each dimension has a default value that is assigned if the event’s textual context does not provide evidence for the assignment of one of the other values.

Knowledge Type (KT): Captures the general information content of the event. Each event is classified as one of the following: *Investigation* (enquiries and examinations), *Observation* (direct experimental observations), *Analysis* (inferences, interpretations and conjectures), *Method* (experimental methods) *Fact* (general facts and well-established knowledge) or *Other* (default: events expressing incomplete information, or whose KT is unclear from the context)

Certainty Level (CL): Encodes the confidence or certainty level ascribed to the event in the given text. We partition the epistemic scale into three distinct levels: *L3* (default: no expression of uncertainty), *L2* (high confidence or slight speculation) and *L1* (low confidence or considerable speculation).

Polarity: Identifies negated events. We define negation as the absence or non-existence of an entity or a process. Possible values are *Positive* (default) and *Negative.*

Manner: Captures information about the rate, level, strength or intensity of the event, using three values: *High* (the event occurs at a high rate or level of intensity), *Low* (the event occurs at a low rate or level of intensity) or *Neutral* (default: no indication of rate/intensity).

Source: Encodes the source of the knowledge being expressed by the event as *Current* (default: the current study) or *Other* (any other source).

Hyper-Dimensions: Correspond to additional information that can be interfered by considering combinations of some of the explicitly annotated dimensions. We have identified two such hyper-dimensions each with binary values (*Yes* or *No*): *New Knowledge* (inferred from *KT*, *Source* and *CL*) and *Hypothesis* (inferred from *KT* and *CL*).

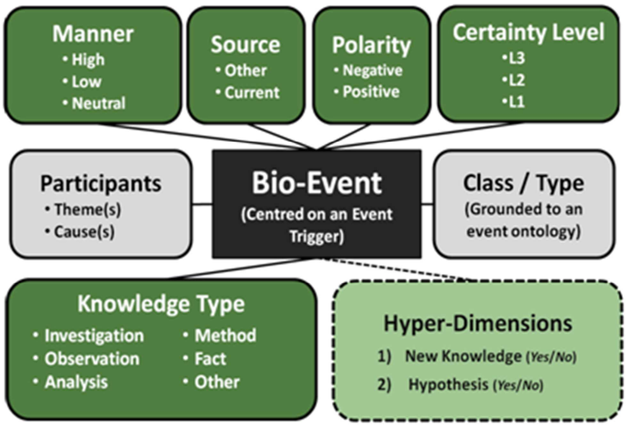

Figure 2: Meta-knowledge annotation scheme

The annotation of the GENIA Event corpus according to this scheme (Thompson et al., 2011) showed that high levels of inter-annotator agreement (between 0.843 and 0.929 Kappa) were achieved by following the 66-page guidelines. Also, given that each of the two annotators had a different background (biology vs. linguistics), it was concluded that specific expertise does not appear necessary to perform meta-knowledge annotation.

In the context of the current case study, it was important to consider whether the meta-knowledge scheme needed to be altered prior to its application to full papers. This consideration is relevant, firstly due to the fact that the scheme was defined only on the basis of examining abstracts, and secondly since previous research into meta-knowledge classification at the sentence or zone level has defined different numbers and types of categories to encode the general information content of the sentence/zone, according to whether abstracts (e.g. (McKnight & Srinivasan, 2003; Ruch et al., 2007; Hirohata et al., 2008)) or full papers (e.g. (Mizuta et al., 2006; Liakata et al., 2010)) are under consideration. For full papers, the number of categories defined can be more than double the number used for abstracts.

The information encoded by the *KT* dimension of the event-based meta-knowledge scheme is somewhat comparable to the above schemes. However, while sentence-based categories are quite strongly tied to structural aspects of the article, with labels such as *background, experiment, conclusion,* etc., the values of the *KT* dimension can be considered more abstract or high level. For example, if several different events occur in *background* and *conclusion* sentences, each event could be assigned a different KT value. That is to say, both sentence types could contain certain events that describe observations, and others that represent analyses. Due to the more abstract level of information encoded by *KT* types, we believe them to be applicable both to abstracts and full papers. They can be considered as complementary to sentence or zone-based schemes, in allowing a finer-grained analysis of the different types of information that can occur within a particular sentence or zone type.

We also envisage that the other dimensions of the scheme do not need to be expanded to allow annotation of full papers, as they all appear to represent general features that can be found in many types of text. For example, the use of three different levels of certainty is in line with an analysis of general characteristics of the English language (Hoye, 1997), rather than being specific to abstracts. The two-way distinctions of the *Polarity* and *Source* dimensions are also observable in any kind of academic writing. Similarly, the information encoded by the *Manner* dimension, whilst more domain specific, should also be applicable to full papers.

The ability to apply the same meta-knowledge scheme to both abstracts and full papers has advantages not only in terms of comparing meta-knowledge characteristics between the two text types, but also in facilitating easy portability/scalability of systems trained to assign meta-knowledge to events either at the abstract or full paper level. In performing meta-knowledge annotation of full papers, careful consideration was given as to whether any aspects of event interpretation were missing from the scheme, or whether there were any events that could not be correctly characterised by the existing categories within the dimensions.

4. Annotation of Full Papers

We have applied our meta-knowledge annotation scheme to four full papers, which had previously been manually annotated with events, according to the GENIA event annotation scheme (Kim et al., 2008). According to the previously proven consistency of the meta-knowledge annotation that can be achieved by following the guidelines (Thompson et al., 2011), regardless of annotator background, the meta-knowledge annotation was carried out manually by one of the authors, who has a background in computational linguistics. All events in the four papers were annotated with meta-knowledge, without any concerns regarding deficiencies in the existing scheme, either in terms of missing dimensions, or missing values in existing dimensions. This suggests that the scheme is fully portable between abstracts and full papers.

Table 1 summarises the distribution of the annotations amongst the different categories for each dimension, and Table 2 shows the most frequent clues for each category and their relative frequencies, i.e., the percentage of events of the specified category in which the clue is annotated. Below, we provide a brief discussion of the results of our new annotation effort. We examine results at the level of the complete papers, and also consider the distributions of annotations within the major sections of the papers, i.e., *Background, Methods, Results, Discussion* and *Conclusion.*

4.1 Knowledge Type (KT)

The most commonly annotated value is *Observation*, constituting just over a third of the total number of events. This is unsurprising, since a large proportion of most biomedical papers would be expected to report on definite experimental observations and results.

Considering individual sections within the full papers, *Observation* events are most prevalent in *Background* (42% of all events in this section type). It may seem surprising that the frequency of *Observation* events in *Background* is greater than in *Results.* However, *Observation* events can refer to previous work as well as current work, and the *Background* section will often refer to findings from a large number of related studies. In the *Results* section, approximately 36% of events describe observations; while in the *Discussion* section, the frequency of such events is even lower (32%). This is to be expected, since greater proportion of this section type would normally be analytical in nature.

Only in a small fraction (12%) of the *Observation* events is the *KT* type determined by the presence of an explicit lexical clue (mostly sensory verbs). In most cases, the tense of the event-trigger and the context of the event (both local and global position within the paper) were found to be important factors.

The second most prevalent category is *Other*. These events generally constitute participants of other events whose *KT* value is *Investigation*, *Analysis* or *Fact*. Out of the context of their parent event, these participant events have no specific *KT* interpretation. No explicit lexical clues were annotated for this category.

A relatively large proportion of events (more than one fifth) belong to the *Analysis* category. This makes sense, given that analytical elements are normally to be found to some extent in most section types in full papers. These include the *Background* section, where such events are most likely to provide overviews or interpretations of previous work, as well the *Results*, *Discussion* and *Conclusions* sections, where analyses, interpretations and conclusions regarding authors' own work most commonly appear. As may be expected, the frequency of *Analysis* events is highest in *Discussion/Conclusion* sections, where they constitute over one quarter (27%) of all events.

An explicit lexical clue was found for each *Analysis* event. The clues comprised verbs, modal auxiliaries and certain adverbs (such as, *thus* and *therefore*).

Almost 6% of the events belong to the *Method* category. Although full papers generally include a fairly large *Methods* section, the small number of events falling into this category is largely because the GENIA event annotation focusses on dynamic relations, i.e., at least one of the biological entities in the relationship is affected, with respect to its properties or its location, in the reported context. This means that descriptions of methods are often less relevant event annotation targets than are events describing observations and analyses.

Our case study suggests that only a small proportion of events in full papers (around 4%) describe factual knowledge. Such events are not evenly distributed throughout papers, and occur most frequently in *Background* (7.5% of all events in this section type), in order to provide context for the new research described in the paper. They can also appear in the *Discussion* section (4.5% of events), where they may be contrasted or compared with the outcomes of the current study. As may be expected, factual knowledge is almost never referred to in the *Results* sections of papers. Similarly to the *Observation* category, most (85%) events from this category did not have an explicit lexical clue.

Dimension	**Category**	**Events**	**Relative Frequency (RF)**
Knowledge Type (KT)	Analysis	381	22.3%
	Investigation	65	3.8%
	Observation	619	36.2%
	Fact	70	4.1%
	Method	100	5.8%
	Other	475	27.8%
Certainty Level (CL)	L1	39	2.3%
	L2	162	9.5%
	L3	1509	88.2%
Polarity	Negative	63	3.7%
	Positive	1647	96.3%
Manner	High	66	3.9%
	Low	15	0.9%
	Neutral	1629	95.3%
Source	Current	1369	80.1%
	Other	341	19.9%
Hyper-Dimensions	New Knowledge	489	28.6%
	Hypothesis	259	15.1%

Table 1: Category distribution

The *Investigation* KT category is the least frequent. The results of our annotation experiment suggest that the *Background* section normally very briefly introduces the subject of investigation (2.5% of events in this section type). A slightly more detailed description of the investigation is then given in the *Results* section (5.4% of all events in this section type). It is also possible that the research goal will be very briefly reintroduced in the

Discussion section of the paper (an average of 1.8% of all events in this section type). All *Investigation* events were accompanied by an explicit lexical clue.

4.2 Certainty Level (CL)

Almost 12% of all events in our full paper sample are expressed with some degree of uncertainty, almost all of which belong to the *KT* type *Analysis*. Taking this into account, the need for this dimension becomes more apparent: whilst under half of *Analysis* events (47%) are stated with no uncertainty, this also means that over a half of these events do express some kind of uncertainty. In fact, 43% of all *Analysis* events are annotated as having slight speculation (*L2*), whilst 10% are reported with greater speculation (*L1*). The marking of uncertainty is sometimes necessary in scientific research literature. Analyses of experimental results may constitute important outcomes, but yet the authors are not confident that their analysis is completely reliable. As stated by Hyland (1996), “Scientists gain credibility by stating the strongest claims they can for their evidence, but they also need to insure against overstatement.” (p. 257). Authors often achieve this by using slight hedging (*L2*). Greater speculation (*L1*) is less common, as credibility is reduced in this case.

Considering individual sections helps to confirm Hyland’s statement. Although the proportion of *Analysis* events that are assigned a *CL* value of *L1* is fairly constant in the *Background, Results* and *Discussion* sections, the proportions of *L2* events have more variation. The relative frequency is lowest in the *Background* sections (36% of *Analysis* events). Since this type of section deals mainly with reporting the work of others, there may be less need to hedge, as it is not the authors’ own credibility at stake. In contrast, the relative frequency of slightly hedged *Analysis* events is noticeably higher in the *Results* and *Discussion* sections (46% and 51%), respectively, where the authors’ own work is the main focus, and hence interpretations and analyses of results are often stated more tentatively.

In terms of clues, modal auxiliaries account for most (70%) of the *L1* events, while the clues for *L2* include both verbs and modals.

4.3 Polarity

Just under 4% of all events are negated. Almost all negated events belong to the KT categories of *Observation* or *Analysis*, which is fairly intuitive. One would not, for example, expect to encounter many cases where *Investigation* or *Method* events are negated. The distributions of negated events vary across different sections of the full papers. The proportions encountered in *Background* and *Discussion* sections are quite similar to each other (around 2% in each section), compared to around 6% of negated events in *Results* sections. Thus, it appears that it is very rare for anything other than positive results to be mentioned in the former two section types. In contrast, when reporting directly on one’s own experimental results, negative results are mentioned more frequently.

Although several negation clues were annotated, the adverbial *not* accounts for over half of negated events.

Dimension	Category	Most Frequent Clues and their RF
Knowledge Type	Analysis	show (16%), demonstrate (14%), indicate (9%), suggest (7%), reveal (5%), can (4%), thus (3%), may (3%)
	Investigation	determine (19%), analyze (15%), elucidate (11%), evaluate (9%), detect (5%), indicate (5%), test (5%), examine (3%), investigate (3%)
	Observation	observe (4%), find (3%), show (1%), document (1%), exhibit (1%)
	Fact	known (6%), well established (3%), well known (2%), fact (2%)
Certainty Level	L1	may (54%), can (15%), possibility (10%), not clear (5%), not understood (5%)
	L2	indicate (22%), can (15%), suggest (11%), ability (6%), able (6%), potential (4%), hypothesize (3%), imply (3%), suspect (3%)
Polarity	Negative	not (57%), no (18%), failure (10%), non (8%), fail (2%), inability (2%)
Manner	High	significantly (17%), well (12%), much (11%), n-fold (9%), strong (9%), strongly (6%), high (3%), higher (3%)
	Low	minimal (13%), little (13%), weak (13%), weaker (13%), n% (7%), less (7%)
Source	Other	Citation (78%), has been (12%), previously (2%), recently (2%)

Table 2: Most frequent clues for each category together with relative frequencies (RF)

4.4 Manner

Almost 5% of all events are expressed with a *Manner* other than *Neutral*. This proportion is fairly constant in the *Background, Results* and *Discussion* sections of the full papers, showing that, although fairly rare, information about the manner of events can be of relevance to the discussion in various different parts of the paper. However, the expression of *High* manner is 4 times more frequent than that of *Low* manner. Similarly to negation, most *High* and *Manner* events belong to *KT* categories of *Observation* or *Analysis*.

Another similar pattern to the *Polarity* dimension is that events with a *Manner* value of *Low* seem to appear with any regularity only in the *Results* sections of the papers, where they appear with just over half the frequency of events whose *Manner* value is *High.* In contrast, the *Low* value was never annotated in the *Background* sections of the papers, and was only annotated for less than 1% of events in the *Discussion* sections. This suggests that events with *Low* manner constitute fairly insignificant information, and are normally mentioned only when reporting experimental results.

Most manner clues are adverbs or adjectives; however numerical values (such as, *n-fold* and *n%*) are also used to express *High* manner.

4.5 Source

Nearly 20% of all events in the full papers belong to the *Other* category. The concentration of such events is highest in the *Background* sections of the papers, where over 40% of the events are attributed to other sources. This is expected, since the *Background* section normally contains the highest concentration of descriptions of previous work. The *Discussion* sections of the papers also have a high (over 25%) concentration of *Other* events, since in this type of section, it is common to compare and contrast the outcomes of the current work with those of previous, related studies. The frequency of *Other* events in the remaining sections is considerably lower. For example, in the *Results* sections of the papers, less than 7% of events are annotated as *Other.* While citations accounted for most of the *Other* events, the use of past perfect tense and explicit markers (such as *previously* and *recently*) also served as clues.

4.6 Hyper-Dimensions

Using the annotations for *KT*, *CL* and *Source* dimensions, we computed the values for the *New Knowledge* and *Hypothesis* dimensions. We found that nearly 29% of all events conveyed new knowledge, and over 15% of all events represented hypotheses. Events conveying new knowledge were predominantly found in the *Results*, *Discussion* and *Conclusion* sections, while hypotheses were found in these sections as well as in the *Background* section. The *Methods* section contained hardly any hypotheses or claims of new knowledge.

5. Comparison with Abstracts

In this section, we compare the distribution of meta-knowledge annotation results obtained in our case study of full papers with those obtained for abstracts, as reported in Thompson et al. (2011). Table 3 shows the difference between the category distributions for full papers and abstracts. Below, we provide a brief discussion of the differences in each dimension.

KT: The biggest difference is seen for the *Method* events, which are more than twice as abundant (in terms of relative frequency) in full papers than in abstracts. This is probably because abstracts tend to focus more on results and their significance, rather than how these results were obtained. As mentioned above, however, the frequency of *Method* events is quite low even for full papers, due to the "dynamic" nature of GENIA events.

A further feature of abstracts is that they tend to contain one or two sentences summarising current knowledge (i.e., well known facts) in the relevant field. Since the average size of abstracts in the GENIA event corpus is 9 to 10 sentences (Kim et al., 2008), the relative frequency of facts in abstracts is quite high (over 8%). This proportion is comparable to the number of factual events in *Background* sections of full papers (over 7% of all events in this section type), where the current state of knowledge is also discussed in some detail. However, as was explained in section 4.1, events describing facts are far scarcer in the other sections of full papers and, given the overall length of papers, the relative frequency of *Fact* events in full papers as a whole is only around half of the frequency in abstracts.

Regarding *Investigation* events, their relative frequency in the *Results* sections of the full papers is comparable to their relative frequency in abstracts (around 5%). However, similarly to the *Fact* category, the extremely rare appearance of *Investigation* events in other sections of full papers means that overall relative frequency in full papers is also much lower than in abstracts.

The relative frequency of *Analysis* events is around 25% higher in full papers than in abstracts. As explained in the previous section, and in contrast to *Fact* and *Investigation* events, *Analysis* events are found with quite high frequency in several sections of full papers. For the *Other* and particularly the *Observation* categories, there is much less variation between the relative frequencies in full papers and abstracts. Thus, clear reporting of experimental observations is equally important throughout both full papers and abstracts.

CL: Owing to the very nature of abstracts, a high proportion of events with no uncertainty is to be expected. As authors aim to "sell" the most positive aspects of their work in abstracts, it makes sense that the majority of analyses should be presented in a confident manner. However, as explained in section 4.2, authors tend to be more cautious while detailing their results and findings in the main body of papers, in order to maintain credibility in case their results are later disproved. The fact that the proportion of slightly hedged *Analysis* events is particularly high in the *Results*, *Discussion* and *Conclusion* sections of full papers, rising as high as 51% in the *Discussion* sections, helps to explain why *L2* events are over 57% more frequent in full papers than in abstracts. The relative frequency of *L1* events is also higher in full papers by about 10%.

Polarity: The relative frequency of negated events is significantly (67%) higher in abstracts than in full papers. This is partly due to the fact that negative results are sometimes more significant than positive results (Knight, 2003), and are therefore, highlighted in the abstracts. In addition, since negated events only appear with any regularity in the *Results* sections of full papers, this helps to explain their lower relative frequency than in abstracts when the complete paper is considered.

Manner: The distribution of *High* and *Neutral* manner is very similar in abstracts and full papers, and the distribution of *Low* manner is exactly same. This follows the same trend described in section 4.4, where it was also noted that the proportions of events with explicit manner markings are also fairly similar across several individual section types within full papers.

Dim.	**Cat.**	**RF (FP)**	**RF (A)**	**Diff. in RF (FP – A)**	**% Change in RF**
KT	Ana.	22.2%	17.8%	4.4%	24.8%
	Inv.	3.8%	5.3%	-1.5%	-39.0%
	Obs.	36.3%	34.7%	1.4%	4.1%
	Fact	4.1%	8.1%	-4.0%	-98.7%
	Meth.	5.8%	2.6%	3.2%	120.8%
	Oth.	27.8%	31.3%	-3.5%	-12.7%
CL	L1	2.3%	2.1%	0.2%	9.7%
	L2	9.5%	6.0%	3.5%	57.6%
	L3	88.2%	91.9%	-3.7%	-4.2%
Pol.	Neg.	3.6%	6.1%	-2.5%	-66.7%
	Pos.	96.4%	93.9%	2.5%	2.6%
Man.	High	3.9%	3.8%	0.1%	2.2%
	Low	0.8%	0.8%	0.0%	0.0%
	Neut.	95.2%	95.3%	-0.1%	-0.1%
Src	Cur.	80.0%	98.5%	-18.5%	-23.1%
	Oth.	20.0%	1.5%	18.5%	1248.6 %
Hyper-D	N.K	28.6%	43.4%	-14.8%	-51.7%
	Hypo.	15.2%	13.4%	1.8%	13.4%

Table 3: Difference between relative frequencies (RF) of categories in full papers (FP) and abstracts (A)

Source: This is the dimension for which the largest difference in category distribution exists between abstracts and full papers. Full papers contain 12.5 times as many *Other* events as abstracts. This is mainly because abstracts are meant to summarise the work carried out in the current study. Furthermore, citations, which are the most common way to denote previous work, are often not allowed in abstracts. In contrast, full papers normally mention related work quite extensively, most notably in *Background* and *Discussion* section.

Hyper-Dimensions: While the relative frequency of *Hypothesis* events is higher in full papers, the proportion of *New Knowledge* events is significantly higher in abstracts. This is mainly because, in abstracts, authors typically include most of new discoveries and results, while only mentioning the main hypotheses.

6. Conclusion

In this article, we have described a case study to investigate the feasibility of applying an event level meta-knowledge annotation scheme (Thompson et al, 2011), whose design was originally guided only by reference to abstracts, to full papers. This is important, given that work on event extraction is gradually being scaled from abstracts to full papers, and also that the automatic recognition of meta-knowledge about events can be highly useful for building more sophisticated IE systems. Our case study involved the annotation of 4 full papers using the meta-knowledge annotation guidelines described in Thompson et al. (2011). The results of the case study strongly suggest that the existing meta-knowledge annotation scheme can be successfully applied to full papers, without any modifications

In order to help to guide the engineering of features for event-based meta-knowledge assignment systems trained on full papers, we conducted an analysis of the meta-knowledge annotations created during our case study. The analysis was concerned not only with the overall distribution of meta-knowledge categories in the full papers, but also with comparisons of the distributions of meta-knowledge categories, both between different sections of the papers, and also with meta-knowledge annotations added to the GENIA Event corpus of MEDLINE abstracts (Thompson et al., 2011). In certain cases, notable differences in the distribution of categories within particular dimensions could be observed both between the different sections of full papers, as well as between full papers and abstracts. This suggests that it may be appropriate to train separate meta-knowledge classifiers for full papers and abstracts. It may also be advantageous to use section-specific classifiers within full papers.

Based upon the demonstrated applicability of the meta-knowledge annotation scheme to full papers, we plan to embark upon a larger annotation effort to enrich all full papers from the BioNLP 2011 GENIA event task with meta-knowledge annotation, in order to increase the amount of annotated data available for training meta-knowledge assignment systems that can operate on full papers. We will also aim to enrich other event-annotated corpora released as part of other tasks in the BioNLP 2011 Shared Task, which include both full papers and abstracts dealing with different domains.

7. 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) and through the MetaNet4U project (ICT PSP Programme, Grant Agreement: No 270893).

8. 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. The Gene Ontology Consortium. *Nature Genetics*, 25(1): 25-29.

Blake, C. (2010). Beyond genes, proteins, and abstracts:

Identifying scientific claims from full-text biomedical articles. *J Biomed Inform*, 43(2): 173-189.

Califf, M.E. and Mooney, R.J. (2003). Bottom-up relational learning of pattern matching rules for information extraction. *The Journal of Machine Learning Research*, 4: 177-210.

Cohen, K.B., Johnson, H.L., Verspoor, K., Roeder, C. and Hunter, L.E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11: 492.

Hirohata, K., Okazaki, N., Ananiadou, S. and Ishizuka, M. (2008). Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing*, pp. 381-388.

Hoye, L. (1997). *Adverbs and modality in English*, Longman.

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

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

Kim, J.-D., Ohta, T., Pyysalo, S., Kano, Y. and Tsujii, J. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In *Proceedings of the Workshop on BioNLP: Shared Task*, pp. 1-9.

Kim, J.D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N. and Tsujii, J. (2011a). Overview of BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 1-6.

Kim, J.D., Wang, Y., Takagi, T. and Yonezawa, A. (2011b). Overview of Genia Event Task in BioNLP Shared Task 2011. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pp. 7-15.

Knight, J. (2003). Null and Void. *Nature*, 422: 554-555.

Liakata, M., Teufel, S., Siddharthan, A. and Batchelor, C. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of LREC 2010*, pp. 2054-2061.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C. and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*, 28(7).

McKnight, L. and Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *AMIA Annu Symp Proc*, pp. 440-4.

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 ACL*, pp. 1017-1024.

Mizuta, Y., Korhonen, A., Mullen, T. and Collier, N. (2006). Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics*, 75(6): 468-487.

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

Oda, K., Kim, J.-D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y. and Tsujii, J.i. (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.

Pyysalo, S., Ohta, T., Cho, H.C., Sullivan, D., Mao, C., Sobral, B., Tsujii, J. and Ananiadou, S. (2010). Towards event extraction from full texts on infectious diseases. In *Proceedings of the BioNLP 2011 Workshop*, pp. 132-140.

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): 195-200.

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): 2086-2093.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1): 233-272.

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

Thompson, P., Nawaz, R., McNaught, J. and Ananiadou, S. (2011). Enriching a biomedical event corpus with meta-knowledge annotation. *BMC Bioinformatics*, 12: 393.

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): i331-i339.

Zweigenbaum, P., Demner-Fushman, D., Yu, H. and Cohen, K.B. (2007). Frontiers of biomedical text mining: current progress. *Brief Bioinform.*, 8(5): 358-375.