Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora

Paul Thompson[1], Philip Cotter[1], Sophia Ananiadou[1], John McNaught[1], Simonetta Montemagni[2], Andrea Trabucco[2], Giulia Venturi[2]

[1]National Centre for Text Mining, University of Manchester, UK

[2]Istituto di Linguistica Computazionale, CNR, Italy

E-mail: {paul.thompson,philip.cotter,sophia.ananiadou,jock.mcnaught}@manchester.ac.uk, {simonetta.montemagni,andrea.trabucco,giulia.venturi}@ilc.cnr.it

Abstract

This paper reports on the design and construction of a bio-event annotated corpus which was developed with a specific view to the acquisition of semantic frames from biomedical corpora. We describe the adopted annotation scheme and the annotation process, which is supported by a dedicated annotation tool. The annotated corpus contains 677 abstracts of biomedical research articles.

1. Introduction

This paper presents a resource being developed in the framework of the European BOOTStrep project (FP6 - 028099) providing a bio-event annotated corpus of biomedical literature. The focus is on gene regulation events in a corpus of MEDLINE abstracts on the subject of E. coli. Events described both by verbs and nominalised verbs, such as *transcription* or *expression*, are annotated. Annotation consists of identifying semantic arguments of the event within the same sentence, and labelling them with event-independent semantic roles and named entity (NE) types. Annotation is carried out using a version of the WordFreak annotation tool (Morton and LaCivita, 2003) that has been customised to the task. The resulting annotated corpus contains 677 abstracts.

The annotated corpus is designed to facilitate acquisition of semantic frames, via the application of a machine-learning algorithm, for inclusion within a large-scale bio-lexicon. This is one of the resources being produced as part of the BOOTStrep project (together with a bio-ontology and a fact store) for use by the biological text mining community. It is planned to release the annotated corpus as a further resource, which could be used for other purposes. For example, whilst the corpus is only partially annotated with NEs, it could be used as part of a training corpus for a NE recognizer that is tuned to the gene regulation domain.

2. The Approach

There exists a number of general language repositories of semantic frames, such as Kipper-Schuler (2005), Palmer et al (2005) and Rupenhoffer et al (2006). These repositories differ according to two main aspects:

a) frame types: how many and which types of event frames are distinguished? Are they applied to individual verbs or classes of verbs? Are they restricted to a predefined set or are they bootstrapped from texts?
b) semantic roles: how fine-grained are the semantic relationships between a predicate and its arguments? Are they frame-dependent or frame-independent?

Annotated corpora are central to the production of such repositories, and much effort has recently been devoted to the development of domain-specific corpora within the biomedical field (see among others Kim and Tsujii, 2006). There now exist several biomedical corpora with event-level annotations and domain-specific semantic frame repositories e.g. Chou et al. (2006), Dolbey et al. (2006), Kim et al. (2008), Kulick et al. (2004), Pyysalo et al. (2007) and Wattarujeekrit et al. (2004).

In a similar way to the open-domain field, the approaches followed in the biomedical field to annotate bio-events can be distinguished with respect to the selection of the frame types and of the semantic roles used for the annotation. Moreover, they can be more finely distinguished according to the *specificity with respect to the domain*, i.e. whether and to what extent annotated event frames are domain-dependent.

Our approach to bio-event annotation combines and builds upon elements of a number of these. In common with Wattarujeekrit et al. (2004), we aim for a set of verb-specific semantic frames, which present the advantage of providing more detailed argument specifications; this is particularly important in the biomedical field, where phrases that identify information such as location, manner, timing and condition are essential for the correct interpretation of events (Tsai et al, 2007).

In contrast to Wattarujeekrit et al. (2004), however, we are using frame-independent semantic roles, which can better capture linguistic generalisations (Cohen and Hunter, 2006). Kim et al. (2008) also use frame-independent roles, although they annotate only a small number of semantic argument types. We build upon this approach by annotating *all* the sublanguage semantic arguments of relevant events, using a larger set of domain-specific semantic roles. In common with many other annotated corpora projects, we also annotate the semantic arguments with NE types; this information can be useful to machine-learning algorithms in helping determine the types of phrase that can fill each semantic role.

Although most of the above resources concentrate on events described by verbs, events described using nominalised verbs, such as *expression, transduction* and *control,* play an important and possibly dominant role in biological texts (Cohen and Hunter, 2006). They are

similar to verbs in both their meaning, and in that they often co-occur with semantic arguments relating to the event, e.g. *the repression of hslJ transcription in Escherichia Coli*. Whilst the corpora described in Kim et al (2008) and Pyysalo et al. (2007) pay some attention to such events, our corpus provides an enhanced treatment of these through the annotation of all specified semantic arguments in a parallel way to verbs. As the semantic argument structures of nominalised verbs may vary from their verbal equivalents, learning separate semantic frame information for them will help to enhance the efficiency of text mining.

Incremental annotation is another qualifying feature of our approach: like Kim et al. (2008) and Kulick et al. (2004), our bio-event information is annotated on top of linguistic annotations which, here, cover morphosyntax and shallow syntax ("chunking"). The advantages of such a choice range from practical ones, i.e. annotated corpora can be produced with much less work, to more substantial ones, i.e. previous levels of annotation can drive the annotation process thus resulting in an increase in efficiency and quality for any new annotation.

3. Annotation Scheme

3.1 Semantic Roles

At the core of our annotation scheme is a set of 12 event-independent semantic roles. These have been defined specifically for the task though the examination of a large number of relevant events within the E. coli corpus. On the one hand, the set of roles used needs to be sufficiently large to be able to characterise *all* instantiated semantic arguments of relevant verbs and nominalised verbs. On the other hand, it is desirable to keep the role set as small and as general as possible. This reduces the burden on annotators, whilst also helping to ensure consistency across the verb frames produced through machine learning .

Event-independent semantic roles have previously been used in large-scale projects involving the production of semantic frames for general language verbs; examples include VerbNet (Kipper-Schuler, 2005) and SIMPLE (Lenci et al., 2000). However, to our knowledge, a set of event-independent roles has not previously been proposed for use in the biological domain. In order to define a set of roles for this domain, we began by examining the sets of roles proposed in VerbNet and SIMPLE, with the assumption that certain semantic roles are common across all domains. This assumption was confirmed by comparison of a number of the roles with examples in our corpus, resulting in our use of roles such as AGENT THEME and SOURCE.

Role Name	Description	Example (**bold** = semantic argument, *italics* = focussed verb)
AGENT	Drives/instigates event	**The narL gene product** *activates* the nitrate reductase operon
THEME	a) Affected by/results from event b)Focus of events describing states	**recA protein** was *induced* by UV radiation The **FNR protein** *resembles* CRP
MANNER	Method/way in which event is carried out	cpxA gene *increases* the levels of csgA transcription by **dephosphorylation** of CpxR
INSTRUMENT	Used to carry out event	EnvZ *functions* through **OmpR** to control NP porin gene expression in Escherichia coli K-12.
LOCATION	Where *complete* event takes place	Phosphorylation of OmpR *modulates* expression of the ompF and ompC genes in **Escherichia coli**
SOURCE	Start point of event	A transducing lambda phage was *isolated* from **a strain** harboring a glpD"lacZ fusion
DESTINATION	End point of event	Transcription of gntT is activated by *binding* of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to **a CRP binding site**
TEMPORAL	Situates event in time or with respect to another event	The Alp protease activity is *detected* in cells **after introduction** of plasmids carrying the alpA gene
CONDITION	Environmental conditions or changes in conditions	Strains carrying a mutation in the crp structural gene fail to *repress* ODC and ADC activities in response to **increased cAMP**
RATE	Change of level or rate	marR mutations *elevated* inaA expression by **10- to 20-fold** over that of the wild-type.
DESCRIPTIVE-AGENT	Descriptive information about AGENT of event	It is likely that HyfR *acts* as **a formate-dependent regulator** of the hyf operon
DESCRIPTIVE-THEME	Descriptive information about THEME of event	The FNR protein *resembles* **CRP**.
PURPOSE	Purpose/reason for event occurring	The fusion strains were *used* **to study** the regulation of the cysB gene by assaying the fused lacZ gene product

Table 1: Semantic roles

Whilst some of the roles proposed for general language use do not seem relevant for events in biological domain, it was equally felt that a number of additional roles needed to be proposed in order to characterise certain arguments of biological events in a satisfactory manner. Thus, our annotation scheme includes the following:

a) two new semantic roles, i.e. CONDITION and MANNER, added as domain-specific;
b) semantic roles particularly important for the precise definition of complex biological relations, even though not necessarily specific to the field, i.e. LOCATION and TEMPORAL (see Tsai et al., 2007);
c) semantic roles widely traceable across all domains.

In addition to the 12 defined roles, a 13th role was additionally made available to annotators, i.e. UNDERSPECIFIED. This role was to be assigned if none of the other roles seemed suitable to characterise the argument, together with a comment describing the perceived role of the argument. The complete set of roles used is summarised in Table 1.

3.2 Named Entity Categorisation

As described above, the semantic roles used to classify arguments within our schema are rather general. Their event-independent nature gives rise to the possibility of a THEME being a processing event, its end product, an ancillary to the process or a component of it, among others. However, such information is encoded into the corpus through named entity tagging. In contrast to other corpora projects, we do not aim to annotate all entities within each abstract. Rather, as our annotation is focused specifically on characterizing the semantic arguments of events, only those entities that occur as semantic arguments of annotated gene regulation events are assigned NE categories.

NE class	Definition
DNA	Entities chiefly composed of nucleic acids and their structural or positional references. This includes the physical structure of all DNA-based entities and the functional roles associated with regions thereof.
PROTEIN	Entities chiefly composed of amino acids and their positional references. This includes the physical structure and functional roles associated with each type.
EXPERIMENTAL	Both physical and methodological entities, either used, consumed or required for a reaction to take place.
ORGANISMS	Entities representing individuals or collections of living things and their component parts.
PROCESSES	Set of *event* classes used to label biological processes

Table 2: Named Entities

Our set of NE tags goes beyond the traditional view of NEs, as labelling is extended to include *events* represented by nominalised verbs (e.g. *repression*). Thus, we created bio-specific NE features, specifically tuned to gene regulation domain. We have defined a set of 61 NE classes, which are divided into four entity-specific super-classes (DNA, PROTEIN, EXPERIMENTAL and ORGANISMS), and one event-specific super-class (PROCESSES). Table 2 provides definitions of these super-classes. The NEs within each of these classes are hierarchically-structured, and annotators were instructed to assign the most specific class possible. The NE categories have subsequently been mapped to the Gene Regulation Ontology (Spelendiani et al, 2007), which has been developed as part of the BOOTStrep project, and integrates parts of other established bio-ontologies, such as the Gene Ontology (Ashburner et al., 2000) and the Sequence Ontology (Eilbeck,2005).

4. Annotation Process

Annotation is carried out on abstracts drawn from a dataset collected by domain experts, containing at least one mention of an E. coli gene or protein name. As the corpus is not specific to the theme of gene regulation, the preliminary step is to verify the thematic content of each abstract prior to annotation.

The annotation of each relevant abstract is primed by automatically marking all instances from a list of 700 biologically relevant verbs provided by domain experts that potentially denote gene regulation events. Annotators perform annotation only on those verbs denoting gene regulation events. The annotated corpus will thus allow us to identify which verbs from the list are most important for the description of gene regulation events, as well as ensuring that the semantic frames learnt for these verbs will be specific to their gene regulation usage.

For each verb that describes a relevant event, annotation proceeds as follows:

1) Mark the semantic arguments of the verb that occur within the same sentence
2) Assign an appropriate semantic role to each argument
3) Label arguments that correspond to biological NEs with appropriate NE types

The same steps are also taken for nominalised verbs, which are annotated only when occurring as semantic arguments of verbs that have already been annotated. The following example illustrates this, where the nominalised verb *expression* is the THEME of the verb *affected*:

Expression of the ompF and ompC genes is *affected* in a reciprocal manner by the osmolarity of the growth medium.

The annotation process is supported by a customised version of WordFreak (Morton & LaCivita, 2003), a Java-based linguistic annotation tool designed to support both human and automatic annotation of linguistic data. Customization of the tool required the definition of a new annotation task specifying the types of annotations the task is based on, namely event frame annotation and NE categorisation. Annotation can be performed against different types of text visualizations, i.e. a tree-like format or continuous text. Given the type of annotation performed, which is constrained to the occurrences of verbs denoting gene regulation events, the tool helps the annotator to find all occurrences of the biologically relevant verbs in the text. Moreover, to increase quality

and consistency of produced annotations, a number of linguistic constraints has been enforced in the tool to help to ensure that valid annotations are produced (e.g. concerning the syntactic category of semantic role fillers). Figure 1 illustrates an example of the annotation displayed in WordFreak format.

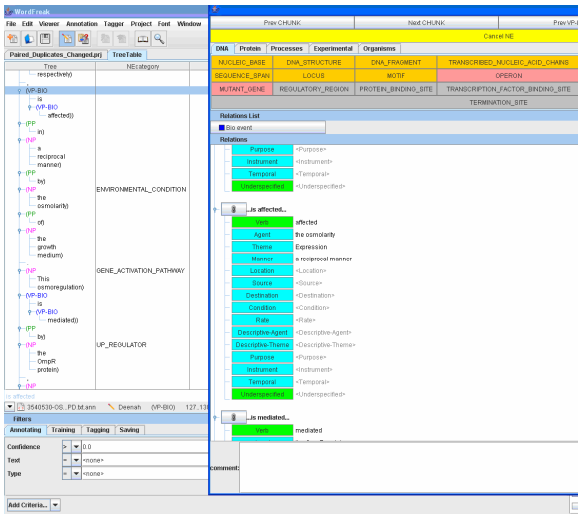

Figure 1 Annotation with WordFreak

4.1 Consistency of Annotations

Annotators were provided with a detailed set of guidelines. These contained, amongst other things, descriptions and several examples of each semantic role, together with explanations of the different NE types. A further set of guidelines aimed to ensure as much consistency as possible amongst the text spans used to represent semantic arguments.

The following sentence serves to illustrate the potentially large variability that could occur in the selection of appropriate text spans to represent semantic arguments:

The Klebsiella rcsA gene ***encoded*** *a polypeptide of 23 kDa*

Without clear guidance, a number of different text spans may be chosen to represent both the AGENT and THEME of *encoded.* For example, the AGENT of *encoded* may be annotated as any of the following spans: *Klebsiella rcsA, Klebsiella rcsA gene,* or *The Klebsiella rcsA gene*. Similar variability could occur for the THEME, which could be *polypeptide*, *a polypeptide* or *a polypeptide of 23 kDa.*

We thus imposed a number of guidelines which would help to constrain annotators' choices regarding which span to annotate. The aim of this was to produce a consistent and "cleanly" annotated corpus, thus enhancing its potential for reusability and for machine-leaning tasks. One way in which consistency is helped is by performing annotation on top of linguistically pre-processed (pos-tagged and chunked) texts[1]. Syntactic chunks were made visible to annotators, and guidelines stated that, in general, semantic arguments should correspond only to complete syntactic chunks. In the example sentence above, *The Klebsiella rcsA gene* is an NP chunk, and so becomes the only choice of span to represent the AGENT of the event.

In addition to consistency, it was additionally felt desirable to ensure that argument text spans are as short as possible. A guideline states that, where the argument corresponds to an entity, only the chunk(s) corresponding to the entity itself, and not any additional information, should be annotated. This guideline should be applied to the THEME in the example sentence above. So, whilst one of the possible options for the span corresponding to this argument is *a polypeptide of 23 kDa*, only the span *a polypeptide* (which is NP chunk) corresponds to the entity itself, with the remainder of the string providing extra information about the entity.

Other guidelines to ensure short spans include the following:

- Spans should correspond to single chunks wherever possible (as long as this is sufficient to characterise the entity).
- Chunks corresponding to names of entities, e.g. *OmpF,* should be favoured over more general characterizations of the entity, e.g *a positive regulator,* in the case that both are present in a sentence.

In addition, there are special guidelines relating to *lists* of entities, i.e.:

- If a semantic argument consists of a list of entities of the same NE type, then the annotated span should only consist of *one* entity in the list.
- If the list contains items of differing NE types, then one item of *each* type should be annotated.

These two guidelines help to ensure that very long semantic argument spans are avoided, whilst also ensuring that the corpus contains as much information as possible about the types of NEs that can occur as arguments of different types of events.

5. Corpus Statistics

Annotation was carried out at the University of Manchester over a period of 3 months by 7 PhD students with experience in gene regulation with native or near-native competence in English. Prior to commencing annotation, the students were required to attend training sessions concerning both the application of the annotation schema and the use of the WordFreak tool. Extensive support was also provided to annotators including a mailing list for query posts and discussions, fortnightly meetings to discuss problems and issues, and regular feedback on the annotations produced.

The annotated corpus is divided into 2 parts, i.e.

1. 597 abstracts, each annotated by a single annotator, containing a total of 3612 events
2. 80 pairs of double-annotated documents, allowing inter-annotator agreement and consistency, and containing 1158 distinct events.

The annotations suggest that a relatively small number of verbs are used to describe gene regulation events. In total, 277 verbs and 135 nominalised verbs were annotated, amongst which 73 verbs and 22 nominalised verbs were annotated 10 times or more. Table 3 shows the 10 most commonly annotated verbs and nominalised verbs in the corpus, together with the number of times they were

[1] Each abstract to be annotated is first pre-processed with the GENIA tagger (Tsuruoka et al, 2005).

annotated, and their type (V=verb, NV= nominalised verb).

The fact that 3 out of these top 10 event focus words are nominalised verbs, including the single most commonly annotated word, i.e. *expression,* provides evidence for the assertion that such words play a dominant role in the description of biomedical events (Cohen & Hunter, 2006), and thus emphasises the importance of annotating semantic frame information for them, in addition to verbs.

Word	Count	Type
expression	409	NV
encode	351	V
transcription	125	NV
bind	110	V
require	100	V
express	93	V
regulate	91	V
synthesis	90	NV
contain	80	V
induce	78	V

Table 3 Most commonly annotated event focus words

5.1 Semantic Roles

The counts of semantic roles assigned to arguments of verbs and nominalised verbs in the single-annotator corpus are shown in Table 4. An interesting point to note is that the UNDERSPECIFIED role was assigned only once during the whole annotation project. It will be recalled that this role was made available to assign to semantic arguments whose role did not seem to be well described by one of the other 12 role labels. This suggests that the originally-defined role set has a sufficient scope to describe the vast majority of semantic arguments of gene regulation events, or at least those occurring within abstracts. Although there is a possibility that annotators may have "pigeonholed" certain arguments into potentially unsuitable categories, our review of a large number of annotated abstracts suggests that this is not a common occurrence.

Role Name	Count
THEME	3353
AGENT	1698
LOCATION	526
CONDITION	239
DESCRIPTIVE-THEME	235
MANNER	223
SOURCE	154
DESTINATION	144
DESCRIPTIVE-AGENT	84
RATE	71
INSTRUMENT	60
PURPOSE	57
TEMPORAL	47
UNDERSPECIFIED	1

Table 4: Semantic role counts

The most commonly occurring roles, by a significant margin, are THEME and AGENT, which is unsurprising given that these represent "core" event information that must be present (or at least implied) for most events to make sense. It may at first seem surprising that the THEME role is assigned over twice as many times as the AGENT role. However, this can be best explained by the high occurrence of events described by nominalised verbs, or verbs in the passive form. In these cases, THEMEs are almost always present, but AGENTs less commonly so.

It is also interesting to note that 3 out of the next 4 most commonly assigned roles, namely LOCATION, CONDITION and MANNER, correspond to those which Tsai et al. (2007) highlighted as vital for the description of biological events. Our results thus confirm their importance, and reinforce the need for both domain-dependent as well as domain-independent roles within our scheme. The least commonly used roles are INSTRUMENT, PURPOSE and TEMPORAL. However, as minimum number of assignments within the corpus is 47, our results suggest that none of our defined semantic roles are redundant.

5.2 Named Entities

The single-annotator corpus contains 5401 named entity annotation. All 61 of the defined categories were assigned at least once, with 50 of them being used 10 or more times. The most frequently assigned categories, together with their counts, are shown in Table 5.

Category	Entity set	Count
GENE	DNA	988
PROTEIN	Protein	602
GENE_ACTIVATION_PATHWAY	Processes	350
ENZYME	Protein	326
PROMOTER	DNA	275
DNA_FRAGMENT	DNA	211
PROKARYOTE_STRAIN	Organisms	191
BIOLOGICAL_PROCESS	Processes	178
OPERON	DNA	155
DNA_STRUCTURE	DNA	130

Table 5 Named Entity counts

The most dominant types of entity are thus DNA-based entities, with proteins also being highly pervasive in the description of gene regulation events. Two of the top ten types correspond to *Processes,* rather than entities. Thus, it is highly common for events themselves to form arguments to verbs, a fact which is backed up by the high occurrence of nominalised verbs. The only set of entities that does not figure in the top 10 is the *Experimental* set. This is perhaps to be expected, given that they are most likely to correspond to less commonly occurring role types, such as CONDITION or INSTRUMENT.

6. Inter-Annotator Agreement

As mentioned above, our corpus annotation phase was subject to rather tight time constraints (around 3 months). Whilst duplicate annotation is vital to ensure consistency amongst annotators, we also wanted to maximize the number of annotated abstracts. The decision was thus taken that only a portion of the abstracts would be duplicate-annotated. The total number of duplicate-annotated abstracts stands at 80 pairs, containing a total of 1158 distinct events. Several

statistics about this corpus are shown in Table 6.
The figures shown in the table are direct agreement rates. Whilst the Kappa statistic is very familiar in calculating inter-annotator agreement, we follow Wilbur et al. (2006) and Pyysalo et al. (2007) in choosing not to use it, because it is not appropriate or possible to calculate it for all of the above statistics. For instance:

1. For some tasks, like annotation of events and arguments spans, deciding how to calculate random agreement is not clear
2. The Kappa statistic assumes that annotation categories are discrete and mutually exclusive. This is not the case for the NE categories, which are hierarchically structured.

Agreement Type	Rate
EVENTS	
Event identification	0.49
ARGUMENTS	
Arg. identification (exact span match)	0.60
Arg. identification (partial span match)	0.73
SEMANTIC ROLES	
Semantic role assignment	0.78
NAMED ENTITIES	
NE identification (exact span match)	0.57
NE identification (partial span match)	0.68
NAMED ENTITY CATEGORIES	
NE cat. assignment (exact)	0.62
NE cat. assignment (including parent)	0.65
NE cat. assignment (including ancestors)	0.73

Table 6 Inter-annotator agreement rates

Table 6 shows that, in terms of events identified, agreement between annotators is reached about half the time. Whilst this figure may seem somewhat low, it will be recalled that annotators had to decide whether each pre-marked verb described an event within the specific range of topics covered by "gene regulation". Several meetings and large amounts of discussion were required in order to reach a consensus on the exact nature of these topics. Thus, particularly towards the start of the annotation phase, annotators tended to either under- or over-annotate the events, which contributed towards the relatively low agreement figure.
Comparisons of other parts of the annotation task show more promising results. For example, annotators agree on the number and location of event arguments in almost three quarters of cases, suggesting that they are fairly reliably able to determine what constitutes a semantic argument of an event. This agreement rate applies only to partial (i.e. overlapping) span matches. If we are stricter, and count only exact span matches, then the agreement rate decreases to 60%. As described above, our annotation guidelines made a considerable effort to enforce consistency of annotated spans. However, our results suggest that there may still be some need to refine the guidelines. A problem here is that the wide range of forms that semantic arguments can take makes the provision of exhaustive examples and general rules for marking consistent span lengths rather difficult.

6.1 Semantic Role Agreement

The highest rate of agreement is for the assignment of semantic roles, at 78%. This suggests that the detailed explanations and examples of the roles provided within the annotator guidelines, together with discussions in the meeting, have equipped annotators with sufficient knowledge to categorise semantic arguments with a relatively high degree of accuracy.
However, as shown in Table 3, the AGENT and THEME roles make up the vast majority of the semantic roles assigned. We also consider these as the most straightforward of the role labels to assign. For this reason, the 78% statistic does not necessarily give a clear indication about how much agreement is reached on the assignment of roles to the less common argument types. We thus calculated agreement rates for the individual semantic roles. These are shown in Table 7, together with a count of the total number of assignments of each role.

Role Name	Count	Agreement
RATE	16	1.00
SOURCE	15	0.93
LOCATION	111	0.90
THEME	975	0.87
AGENT	434	0.85
CONDITION	40	0.80
TEMPORAL	8	0.75
MANNER	59	0.71
PURPOSE	16	0.63
INSTRUMENT	7	0.57
DESTINATION	36	0.44
DESCRIPTIVE-THEME	104	0.46
DESCRIPTIVE-AGENT	35	0.23

Table 7 Agreement rates amongst semantic roles

Whilst the agreement rates for the less commonly occurring roles may not be fully reliable due to their sparseness,, the table shows that the agreement rates for many of the most frequently occurring roles (i.e. AGENT, THEME, LOCATION, MANNER and CONDITION) lie between 70% and 90%, and thus seem acceptably high.
The three roles with the lowest agreement rates all have a reasonable number of occurrences (particularly DESCRIPTIVE-THEME), suggesting that these statistics are fairly accurate. In order to try to understand these low agreement rates, we first calculated the types of role disagreements that occur in the corpus. The most common of these are shown in Table 8. In the table, the columns "Annotator #1 role" and "Annotator #2 role" correspond to the different roles assigned by the 2 annotators.
The table shows that the 4 most common disagreements between role assignments all involve the THEME role. Closer examination of the annotated events corresponding to the first 3 types of disagreement reveals that they mainly concern 3 verbs, namely *encode*, *code* and *bind*.
A typical sentence in which disagreement occurs is the following:

***malS**, the gene **encoding** **the periplasmic alpha-amylase**, is under the regulatory control of the MalT protein.*

For such sentences, there are three common patterns of role assignment for the semantic arguments of *encode*, as

shown in Table 9.

Annotator#1 role	Annotator #2 role	Count
THEME	AGENT	52
THEME	DESCRIPTIVE-THEME	38
THEME	DESCRIPTIVE-AGENT	20
THEME	DESTINATION	11
DESCRIPTIVE-THEME	DESCRIPTIVE-AGENT	6
THEME	MANNER	6
PURPOSE	AGENT	6
DESCRIPTIVE-THEME	AGENT	5
LOCATION	DESTINATION	5

Table 8 Most common role disagreements

malS	***The periplasmic alpha-amylase***
AGENT	THEME
THEME	DESCRIPTIVE-THEME
AGENT	DESCRIPTIVE-AGENT

Table 9 Semantic role assignment patterns for *encode*

The choice of pattern corresponds to the annotator's interpretation of the event's semantics. The AGENT/THEME pattern is most appropriate when the verb describes an action of some kind, whilst the THEME/DESCRIPTIVE-THEME pattern is more suitable when the verb describes a state (i.e. when there no action involved and hence no AGENT). Indeed, the difficulty in annotating *encode, bind* and *code* was discussed during the regular meetings, where it was suggested that their interpretation can vary according to context, and hence both of these patterns may be appropriate for different occurrences of the verbs. However, the fact that there is a fair amount of disagreement of patterns used for particular instances of these verbs suggests the correct interpretation is not always easy to determine, even for domain experts.

For *encode,* a third pattern is observable, namely assigning AGENT to the logical subject of the verb, and DESCRIPTIVE-AGENT to the object. This interpretation suggests that action is involved in the event, but that the subject provides descriptive information about the agent, rather than corresponding to something directly affected by the event. Closer examination showed that this pattern was only used by one annotator. However, it emphasizes the difficulty in correctly categorizing the semantic arguments of this verb in particular.

For the confusions involving the DESTINATION role, the main verb involved is *bind*, as in the following sentence:

In contrast, the ***OmpR2 protein bound*** *preferentially to* ***the ompF promoter***.

The problem again seems to be one of varying interpretation of the *binding* event. One possibility is that an AGENT (the subject of the verb) actively binds to a DESTINATION. Another interpretation is that there is no explicit AGENT, and that the entities corresponding to the semantic arguments just happen to bind together, in which case they are both annotated as THEMEs. Table 7 shows that a second type of confusion is between DESTINATION and LOCATION. This is an understandable confusion, as both roles correspond to locative information.

6.2. Named Entity Agreement

Agreement rates for the identification of NEs are comparable to those achieved for semantic argument identification. As with argument identification, figures in Table 6 are shown for both exact and partial span matches. NE annotations are only assigned to text spans that correspond to biological NEs or processes, and in many cases, the NE only forms part of the complete semantic argument. Thus, there is scope for disagreement amongst annotators regarding which spans to tag as named entities, and the exact extent of the span to annotate. A further potential problem concerned the annotation of NEs within WordFreak. In their normal method of working, annotators had to switch between different views of the text to annotate semantic roles and NEs. Thus, it is possible that annotators sometimes forgot to assign NE categories.

For those NEs whose identification was agreed upon, we additionally calculated statistics regarding the agreement of categories assigned to them. Whilst the agreement rate for exact category matches is not particularly high (62%), we tried relaxing the matching conditions by taking into consideration the hierarchical structure of the NE categories. By counting as a match the cases where the category assigned by one annotator was the parent of the category assigned by the other annotator, the agreement rate was slightly increased. A more marked increase was achieved if ancestor categories, as well as immediate parents, also counted as matches.

As there is such a large number of NE categories (i.e. 61), deciding the most appropriate category is often quite a complex task, as verified by annotators' comments in the regular meetings. Therefore, high rates of agreement on the exact category to assign may be difficult to achieve. However, the hierarchical structure means that it would be possible to use a smaller set of categories by mapping the specific categories to more general ones.

7. Conclusion

We have presented the design of a scheme for annotating events in biomedical texts, and have reported on its application to gene regulation events within MEDLINE abstracts. To our knowledge, our scheme is unique within the biomedical field in that it involves the use of a fixed set of event-independent semantic roles, made up of both domain-specific and domain independent roles, to characterize *all* instantiated arguments of events. A further feature of our approach is that events described by nominalised verbs are annotated for their event structure in a parallel way to verbs. In additional, a rich set of named entity labels is used to characterize the majority of semantic arguments.

Our results suggest that our scheme is well suited to describing events within the domain. On the one hand, all roles within in the scheme were assigned a sufficient number of times provide evidence of their usefulness. On the other hand, it seems that our proposed role set is sufficiently general and wide-ranging to characterize the

vast majority of semantic arguments of gene regulation events. Furthermore, our inter-annotator agreement rates suggest that semantic arguments can be identified and categorized fairly consistently by different annotators. Where disagreements did occur, these were found to be concentrated on a relatively small number of verbs, with fairly regular alternations of role assignment patterns.

Examination of our annotated corpus has, however, identified a number of problematic areas. These include the choice of which verbs to annotate as gene regulation events, which exact text spans to annotate to represent semantic arguments, and the choice of the most appropriate named entity categories. Whilst guidance relating to all of these is provided in the annotation guidelines, the lower inter-annotator agreement rates for these annotation subtasks suggest that the guidelines may benefit from some revision prior to carrying out any subsequent annotation based on this scheme. Furthermore, the higher agreement rates achieved when considering higher-level NE categories suggest that the current set of categories may be too fine-grained.

Work is currently being carried out on the extraction of bio-event frames from the corpus. Results from this will provide further evidence regarding the effectiveness and adequacy of the annotation scheme for its main intended purpose.

In the future, we also plan to apply our scheme to a wider range of text types within the domain. Firstly, we would like to verify whether our scheme is suitable for application to a wider range of event types within the biology/biomedical field. In addition, we would like to discover how well our scheme applies to events that occur in full-text articles, where the structure of the language used in somewhat different from that used within abstracts.

8. Acknowledgements

This work described in this paper has been funded by the European BOOTStrep project (FP6 - 028099).

9. 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. & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetetics,* 25, 25--29.

Chou, W.C., Tsai, R.T.H., Su, Y.S., Ku, W., Sung, T.Y. & Hsu, W.L. (2006). A Semi-Automatic Method for Annotating a Biomedical Proposition Bank. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006,* pp 5--12.

Cohen, K.B & Hunter, L. (2006). A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics* 7 (Suppl. 3), S5.

Dolbey, A., Ellsworth, M. & Scheffczykx, J. (2006). BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In O. Bodenreider (Ed.), *Proceedings of KR-MED,* pp 87--94.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. & Ashburner, M. (2005) The Sequence Ontology: A tool for the unification of genome annotations. *Genome Biology* 6:R44

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

Kim, J-D & Tsujii, J. (2006), Corpora and their Annotation. In. S. Ananiadou & J. McNaught (Eds.), Text Mining for Biology and Biomedicine, London: Artech House, pp 179--212.

Kipper-Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon. PhD. Thesis. Computer and Information Science Dept., University of Pennsylvania. Philadelphia, PA.

Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A. & Ungar, L. (2004) Integrated Annotation for Biomedical Information Extraction. In *HLT-NAACL 2004 Workshop: BioLink 2004, Linking Biological Literature, Ontologies and Databases*, pp 61--68.

Lenci, A., Busa, F., Ruimy, N., Gola, E., Monachini, M., Calzolari, N., Zampolli A., Guimier, E., Recourcé, G. Humphreys, L., Von Rekovsky, U., Ogonowski, A., McCauley, C., Peters, W., Peters, I, Gaizauskas, R & Villegas, M. (2000). SIMPLE Linguistic Specifications LE-SIMPLE (LE4-8346), Deliverable D2.1 & D2.2. ILC and University of Pisa, Pisa.

Morton T. & LaCivita J. (2003). Word-Freak: an open tool for linguistic annotation. In *Proceedings of HLT/NAACL-2003*, pp 17--18.

Palmer M., Kingsbury P. & Gildea D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), pp 71--106.

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

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. & Scheffczyk, J. (2006). FrameNet II: Extended Theory and Practice Available online at http://framenet.icsi.berkeley.edu/

Splendiani, A., Beisswanger, E., Kim, J-J., Lee, V. , Dameron, O. & Rebholz-Schuhmann, D. (2007). Bio-Ontologies in the context of the BOOTStrep project. *Bio-Ontologies SIG Workshop*, Vienna.

Tsai R.T.H, Chou W.C., Su Y.S., Lin Y.C., Sung C.L., Dai H.J, Yeh I.T.H., Ku W, Sung T.Y & Hsu W.L. (2007). BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features, *BMC Bioinformatics* 8:325

Tsuruoka, Y., Tateishi, Y., Kim, J-D., Ohta, T., McNaught, J., Ananiadou, S. & Tsujii, J. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text, In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp 382--392.

Wattarujeekrit, T., Shah, P. & Collier, N. (2004). PASBio: predicate-argument structures for event extraction in molecular biology, *BMC Bioinformatics* 5:155