Improving Search through Event-based Biomedical Text Mining

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

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

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

ABSTRACT

Recently, there has been a major focus on event extraction for biomedical applications. In this paper, we focus on search, highlight some of the drawbacks of popular search methods, and show how event extraction and associated technologies, e.g., named entity recognition, can help to improve the efficiency of search. We also explore how event extraction can be enhanced through a new type of annotation, i.e. meta-knowledge annotation, which can facilitate the extraction of high-level information relating to the intended interpretation of events, e.g. whether they represent a hypothesis, a claim, a belief, an opinion, a well established fact, a tentative or more confident analysis of experimental results, etc.

1. BACKGROUND

The amount of biomedical literature is increasing at a rapid rate, with the size of PubMed increasing at the rate of approximately 2 papers per minute [17]. As a result, it is becoming increasingly difficult for biologists to locate information relevant to their research contained within textual documents.

The goal of searching the literature is to find relevant pieces of *knowledge* (e.g., biological processes).. Suppose that a biologist is interested in discovering which proteins are *positively regulated* by the protein IL-2. An example of the type of sentence she wishes to locate is the following:

IL-2 activates p21ras proteins in normal human T lymphocytes.

This sentence allows the biologist to discover that *p21ras proteins* are one type of protein to satisfy her query. To locate such sentences of interest using an ordinary search engine, the biologist may enter the search terms *IL-2* and *activates*. Such a query will, however, return a large number of documents. On the one hand, many of the documents are likely to be irrelevant to the user's query. On the other hand, the query also has a high probability of missing documents that are relevant to the user's requirements. The reasons for these problems include:

First International AMICUS Workshop, October 21, 2010, Vienna, Austria.

- **Specificity of the query –** The query will over-generate (i.e. return too many documents) because it cannot convey the biologist's specific requirements. In terms of semantics, the user wishes to retrieve documents conveying sentences that describe positive regulations, where *IL-2* is the instigator of the regulation. Such knowledge is normally expressed according to the syntactic structure of the sentence. In the case of the verb *activate,* for example, the instigator corresponds to the grammatical subject. Hence, the user would only be interested in sentences where *IL-2* is the grammatical subject of *activate.* In ordinary search engines, however, it is not possible to specify how search terms should be related to each other. This is because the search engine simply sees documents as "bags of words" that have no internal structure.
- **Term variation/ambiguity –** Terms in biomedicine are complex; they include an enormous amount of synonyms and different variant term forms are used in the literature [21]. For example, *IL-2* can also appear without a hyphen (*IL 2)*. These terms are themselves acronyms for the longer forms *interleukin 2* and *interleukin-2.* The term *T-cell growth factor* can be considered as a synonym of *interleukin-2*, but this also has its own variant forms (e.g. *TCGF*). Thus, using *IL-2* as a search term without considering its variants would result in many relevant documents being overlooked. For query formulation, we need techniques which include term variation. In addition, many terms and their variants are ambiguous, as they share lexical representations either with common English words (e.g. an, by, cat, can) which also denote gene/protein names, or with other biomedical terms [7]. A further issue concerns terms that can also be ambiguous between biological and general English senses, e.g., the proteins named *cat* and *met.* Using such common words as search terms will return many more irrelevant documents than relevant ones, if only documents containing the protein names are sought.
- **Different ways of expressing knowledge –** The verb *activate* is only one of the ways in which positive regulations can be expressed in texts. Other verbs could also be used, e.g., *stimulate* or *affect*, whilst nouns (nominalised verbs) convey a similar meaning, e.g., *activation, effect, stimulation* could also be substituted. As with term variation, it can be difficult to enumerate all the ways in which a particular type of biological process can be expressed. However, if they are not accounted for in a query, then relevant documents may be missed.

The application of text mining methods[2, 3, 52] can provide solutions to the problems outlined above, and can thus contribute to more efficient and effective search solutions for biologists. Text mining techniques can help to ensure that a greater number of relevant search results are obtained, whilst helping to exclude those results that are irrelevant to the user's query.

The remainder of this paper will focus on a number of these methods, and explain how they can improve search results. We firstly look at named entity recognition (NER)[1], and we will examine search engines which offer advanced search capabilities based on semantic metadata derived from named entities and relations. .

NER is one of the technologies that is required to perform extraction and querying of *events* [4]. Events are structured, semantic representations of pieces of knowledge contained within text. We focus on relations [39] within the biomedical domain, such as descriptions of positive regulation, transcription, gene expression, [6] etc. Through a combination of a number of techniques, such as deep syntactic parsing [26] and NER, event extraction automatically locates events in texts and identifies their individual participants, e.g., the instigator of the event, the location of the event, etc. This allows users to formulate more structured queries that are better related to their actual needs, e.g., the biologist can request the system to return only those documents that mention a positive regulation event, where *IL-2* is the instigator. The system will retrieve only those documents where the specified relationship exists between the search terms.

Following a detailed examination of how events are extracted and can be used in advanced searching, we conclude by examining a new direction of research, i.e., how interpretative information about events can be captured automatically to further enhance event-based searching. For example, an event may represent a generally accepted fact, a hypothesis, an experimental observation, a tentative analysis of experimental results, etc. These different types of information could be important to the biologist, e.g., some biologists may be interested only in retrieving events that correspond to "reliable" pieces of knowledge, rather than hypotheses or hedged interpretations. This is particularly important for maintaining curated databases of biological knowledge [5]. Other biologists may be interested in matching up hypotheses with proven results.

2. NAMED ENTITY RECOGNITION

A large amount of work has been carried out on the automatic recognition of biologically relevant NEs in texts [39]. This activity is important for a number of reasons:

- It can resolve ambiguities between words used in general language and those that represent biomedical entities (e.g. *cat*)
- It can facilitate mapping terms found in texts to entries in curated biological databases, such as UniProt[46] and Entrez Gene. (http://www.ncbi.nlm.nih.gov/gene), or resources such as the BioThesaurus [22, 51]. This can facilitate direct access from search results to detailed information about biological entities found within the database.
- Automatic highlighting of different recognised NEs in retrieved documents can facilitate a quick skimming of the main content of the document.
- NEs map to event participants, e.g. a cell group (filament) participates in a localisation event in angiogenesis. Hence, NER is a necessary pre-processing step in event extraction.

Given the huge amount of variation of terms in biomedical text [38], work has also been carried out on recognising and resolving several types of variations. Although some variations are listed in curated databases, many are missing. However, it is important that even unlisted term variations can be resolved to the entity that they describe via term normalisation [45], to facilitate their linking with the correct biological database entry. Work on term normalisation was recognised as an important task through its incorporation as a BioCreative task [15]; term normalisation also includes the recognition of acronyms [30], and the use of soft-string matching techniques [45] that recognise new variants of known terms.

2.1 Search Engines Incorporating NER

Conventional information retrieval technology, while very good at handling large scale collections, remains at a rough granular level. Semantic metadata generated from named entities (NEs) (e.g. PROTEIN:IL-1, ORGAN:brain) are helpful for increasing granularity of document search. However, conventional information retrieval systems do not allow users to specify in their query the semantic metadata they are interested in. Lack of such functionality restricts users' potential to search and retrieve documents based on their personal and social profiles. Metadata are critical to enhancing user experience of search: for example, they can support improved personalized search. The richer the metadata, and the more they are linked in to other resources of different types (including e.g., experimental data), the better the search experience. NLM's PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) is a primary search facility for biomedical literature. Other search interfaces based on PubMed focus on ranking citations [40], incorporating external web services [10] or using Web 2.0 technologies [27] to enhance user experience in searching. In CiteXplore (2009), text mining results are included in the search based on Whatizit, EBIMed [34, 35] and iHOP [16] such as protein/gene annotations and protein-protein interactions. KLEIO [29] is an intelligent search engine which provides interactive faceted semantic search over MEDLINE based on NEs. Faceted navigation is proposed as a component of a superior interface for searching metadata in a more interactive and flexible manner [12, 13] and has become adopted on several web sites (e.g., http://express.ebay.com). One of the main criticisms of the conventional search systems freely available is that search queries are effective only when well crafted [11]. In KLEIO, the user can select, using an interactive faceted query builder, the types of semantic queries of interest, suggested by the system. KLEIO delivers rapid responses, based on pre-indexed NEs linked with term variants, includes query expansion with dynamic reclassification of results, linking of all NEs with unique identifiers from a variety of databases, and highlighting of the retrieved documents with the NEs identified. It also integrates term normalisation[44] and links to curated databases [45] to facilitate more focussed searches. KLEIO permits operators corresponding to different types of NEs as an integral part of the user's search. Thus, it is possible to specify

PROTEIN:cat, to ensure that only documents containing an instance of the word *cat* that have been recognised as an NE of type *protein* will be returned by the search. This can considerably reduce the number of documents returned: performing a search only for *cat* in KLEIO retrieves 67159 abstracts, whilst the search *PROTEIN:cat* reduces this number to only 195 abstracts. KLEIO also highlights instances of NEs found in documents, identifies and resolves acronyms to their full forms (through the integration of AcroMine [31] and additionally allows searches to be expanded to include variant terms through soft-string matching and database linking.

Whilst these search engines provide improvements and useful functionalities not found in traditional search engines, they still do not facilitate the precise querying of events. Searches carried out in KLEIO work along similar principles to traditional search engines. So, although NE operators can help to refine the scope of the search, it is still not possible to specify relationships between these terms.

3. EVENT EXTRACTION

Although NER and normalization have been helpful for increasing the specificity of document searches in TM systems such as KLEIO (www.nactem.ac.uk/software/kleio/) and for significantly reducing errors compared with simple keyword-based retrieval, other search systems, such as FACTA [15], use co-occurrence statistics for normalized names in text to enhance the discovery of hidden associations among entities. However, textual co-occurrence of entities does not necessarily indicate meaningful relationships. For this, more advanced analytical methods are necessary, namely methods that undertake deeper semantic analysis. To achieve this aim, techniques have been developed that automatically extract biological events[4]. Recognition of events and their participants is often reliant on a structural analysis of the sentence containing the event [26]. As described above, event participants are often organised around either a verb or a nominalised verb (e.g., *activation*). In this case, event participants normally constitute some or all of the words and phrases that are syntactically (i.e., structurally) related to the verb or nominalised verb in question, especially if these phrases constitute or contain NEs. Thus, full or partial syntactic parsing of text, which provides a structural analysis of a sentence, is normally one of the necessary steps of event extraction, in addition to NER.

MEDIE [25] is a search engine for MEDLINE abstracts that combines many of the features found in KLEIO (e.g. recognition of NEs, linking with databases and ontology and subsequent identification of term variants) with a further annotation processing step that involves the structural analysis of the abstracts.. The NLP modules used for the annotation include (but not limited to) a deep syntactic analyzer, an event expression recognizer and a term recognizer. The syntactic analyzer, Enju parser [26], produces a syntactic and semantic analysis of the text, based on the linguistic formalism of HPSG. A relational concept, such as 'protein A activates protein B', can be precisely described as a query which specifies the semantic structure given by the Enju parser as a set of constraints. This is the main strength of MEDIE compared to other publicly available TM modules which use Boolean formulas of keywords or concepts for query formulation. Boolean formulas basically specify co-occurrence of concepts or words as a constraint for retrieval. One can only specify co-occurrence of protein A, protein B and the verb 'to activate' in the same textual unit (usually an abstract) as a constraint, which results in a large number of false positives. Units of retrieval in MEDIE are finer than those in other TM modules. They can be individual sentences in abstracts, or even phrases. MEDIE accepts a search query through an API, in addition to an interactive search UI. The API takes a tuple of <subject, verb, object> as the input, which describes a biological event/relation, such as <p53, activate, beta-4>, and returns a set of articles in which the event/relation is mentioned.

A further feature of MEDIE is that search results are not only limited to those where the value of the *verb* slot corresponds to *activate*. Rather, through reference to the Gene Ontology [5], events centred on other verbs such as *stimulate, induce, augment, enhance*, etc. are also retrieved by the search. Due to the deep parsing technology used, the template to be completed by the user can be considered as an abstract representation of the way that the events actually manifest themselves in the text. For example, *IL-2* would also be identified as the subject of *activate* in a passive sentence such as *p21ras proteins are activated by IL-2*. Additional features include the ability to specify the types of sentences in which the search should be conducted, e.g. *conclusion, method, result,* etc.

3.1 Additional Event Participants and Semantic Representations

Despite its clear advantages over a traditional search engine, MEDIE still presents some limitations. Firstly, events often have more than two participants. In biomedical texts in particular, information corresponding to locations, time, environmental conditions and manner is considered to be highly important to their correct interpretation [43]. It would be useful to be able to identify these types of information separately, in order to allow restrictions to be placed on their values as part of a search, and also to allow them to be displayed as part of the search results.

Secondly, the search template to be filled is closely tied to the syntactic structure of the text. Searching using a higher level semantic representation would, however, be preferable. In the search problem introduced earlier, we wanted to find events where *IL-2* is the instigator of the positive regulation event. Instigators can also be identified in many other types of events, and so we can assign this type of event participant a general *semantic role* label that will be common across many different types of events, i.e., AGENT. Likewise, most events also specify as a participant the thing that is affected by or during the event, e.g., the protein undergoing positive regulation. The semantic role label that is normally used for such participants is THEME.

AGENT and THEME frequently correspond to the grammatical subject and object of verb, respectively. However, this is not always the case, and there is no consistent correspondence between grammatical positions and semantic roles. Thus, using AGENT and THEME rather than subject and object would allow the event search template to be more general and less tied to syntactic structure of the text. A semantic approach is even more desirable if additional participants (e.g. location, environmental conditions, etc.) are specified as part of the search. Several of these participant types are specified through syntactically similar means, i.e., through the use of prepositional or adverbial phrases [47]. Consider the following example:

A promoter has been identified that directs relA gene transcription towards the pyrG gene in a counterclockwise direction on the E. Coli chromosome

In addition to a subject and an object, the verb *directs* occurs with 3 arguments corresponding to prepositional phrases, each of which corresponds to a different semantic role (namely DESTINATION, MANNER and LOCATION). Although the different prepositions can be used to help in distinguishing between different semantic roles, there is not a one-to-one mapping between prepositions and semantic roles. For example, *in* is used in the above example to introduce a MANNER, but it could equally introduce a LOCATION, e.g., *in E. coli.* By allowing search criteria to include semantic role labels such as LOCATION and MANNER, the user could specify semantically precise search criteria without having to worry about their exact form in the text (e.g., which preposition is used, etc.)

Considering the above, the type of semantic representation that would ideally be produced for the positive regulation event in the sentence *IL-2 activates p21ras proteins in normal human T lymphocytes* is as follows:

EVENT_TYPE: *positive_regulation*

AGENT: *IL-2:PROTEIN*

THEME: *p21ras proteins:PROTEIN*

LOCATION: *in normal human T lymphocytes:CELL*

In the above representation, the event has been assigned a semantic type, i.e. *positive_regulation*. This event type is a label selected from a fixed, ontological set of relevant event types. Others would include *binding, gene_expression,* etc. Additionally, each participant of the event has been separately identified and assigned a semantic role. The NEs within each participant have also been identified and assigned appropriate NE types. Such a representation allows structured searches to be performed with the following types of criteria:

- Ontological classes of events as an alternative to specifying particular verbs.
- Specifications of the participants that should be present in the event (in terms of semantic roles)
- Restrictions on the values of particular participants. These restrictions could take the form of actual entities (e.g. *NF-kappa B*), NE classes (e.g. *PROTEIN*), or a combination of both, in a similar way to KLEIO.

The main challenges of producing a system that can produce such a representation of events are the following:

1) How each type of event manifests itself in the text - they are often organised around a particular set of verbs and nominalised verbs.
2) How syntactically related arguments of the verb/nominalised verb map to semantic roles.

Although grammatical parsers such as Enju [24] have reached an appropriately mature level, the same cannot be said for semantic parsers. This means that the mapping between syntactic and semantic representations is not straightforward. This is complicated by the fact that different verbs behave in idiosyncratic ways, with different numbers of syntactic arguments, which can map in different ways to semantic roles.

3.2 Annotated Corpora

The approaches used to map between the syntactic and semantic levels can vary in a number of ways, both in the method used (rule based vs. machine learning approaches) and the types of external resources employed (either lexical or ontological).

Whether a rule-based or machine learning approach is taken, annotated corpora of events are a vital resource for the development of event extraction systems. These corpora provide direct evidence of how events manifest themselves in texts, and as such, they can be used in both the development/training of event extraction systems, as well as in the evaluation of the performance of such systems, by acting as a "gold standard" [14].

The various event corpora that have been produced in the biomedical field generally have in common that they identify an "anchor" expression (e.g., a verb or nominalised verb) around which the event is organised. Event participants are then individually identified and linked to this anchor expression. BioInfer [33] (1100 sentences) concentrates on identifying core participants (i.e., agent or theme type roles), although they are not labelled with semantic roles. Events are, however classified according to an ontology, thus facilitating the discovery of the ways in which different types of events can be expressed in the text

A similar type of event classification is carried out in the GENIA event corpus [19]. This is a larger corpus, consisting of 1000 MEDLINE abstracts, containing over 9000 sentences. In the GENIA corpus, event participants are classified using semantic roles. Although the focus is on identifying the THEME and CAUSE (similar to AGENT) roles, 3 other types of event participants are also identified and labelled, i.e., location, time and experimental methods.

GREC [41] is a smaller corpus of 240 abstracts, but with a richer type of semantic annotation that is focussed on gene regulation and expression events that are described by verbs and nominalised verbs. For each event, all participants (arguments) in the same sentence are identified and assigned a semantic role from a rich set of 13 roles, tailored to biomedical research articles.

Although GREC is a relatively small annotated corpus compared to GENIA, a recent study [23] has shown that combining smaller, richly annotated corpora with larger corpora that are slightly poorer in information content can help to improve the performance of event extraction system. Whilst the benefits of combining disparate sources in machine learning are well known, this idea is especially attractive, given that the production of large, richly annotated corpora can be very time-consuming.

3.3 Lexical Resources as an Aid to Event Extraction: the BioLexicon

Whilst event extraction systems can be trained based on annotated corpora alone, the use of a computational lexical resource with information about the syntactic and semantic behavioural of verbs within the domain can be used to boost the performance of such systems.

Although syntactic parsers can be used to identify the core arguments of a verb, the idiosyncratic behaviours of individual

verbs within the biomedical domain means that determining which of the modifier phrases (i.e., prepositional and adverbial phrases) should be treated as arguments of the verb (and hence as event participants) can be problematic. As mentioned above, such phrases can correspond to vital pieces of information about the event, such as locations, manners, conditions, etc. By accessing information about typical patterns of syntactic behaviour for individual verbs, we can expect that the event extraction system will do a better job of determining which phrases in the sentence correspond to event participants. This is particularly important in the case of sentences that contain multiple events, in order to determine which phrases are participants of which event. In the following example, each underlined verb corresponds to a different event, with different sets of participants:

IHF may inhibit ompF transcription by altering how OmpR interacts with the ompF promoter

After the event participants have been identified, lexical resources can also help by providing verb-specific mappings from the syntactic arguments to appropriate semantic roles.

Extensive computational lexicons have been constructed for use in processing general English texts (e.g. [20, 32, 36]), but these are not suitable for processing biomedical texts for a number of reasons. Firstly, there are many verbs that are domain specific (e.g., *methylate, phospholylate,* etc.). Other verbs appear in both domains (e.g. *activate*), but are likely to have different behaviours. In general, verbs in the general language domain have fewer arguments, largely due to the fact that modifier phrases are often considered to be less tightly associated with the verb than in biomedical texts.

Until recently, an extensive computational lexical resource comparable to those produced for the general English language domain was not available for the biomedical domain. Resources that had been built were either very small [9, 49] or did not contain semantic information [8].

The BioLexicon [37] is a reusable lexical and conceptual resource suitable for advanced biomedical text mining. One of its defining features is to include a wide range of biomedical terms and variant forms, to facilitate accurate NER in a range of biomedical text mining applications. Integration of the BioLexicon within a biomedical search engine would, for example, allow search terms entered in user's queries to be expanded with their known variants, to ensure that a greater number of relevant documents are retrieved

The BioLexicon gathers together terms and their variants from a number of different curated databases and ontologies into a single unified resource. The original database identifiers for each term are preserved, in order to facilitate linking to information in the source databases. A particular innovation of the BioLexicon is the application of text mining methods to recognise new variants of gene and protein names that appear in biomedical abstracts (MEDLINE) but not in existing databases. Genes and proteins tend to exhibit the greatest amount of variation amongst all types of biomedical entities. Application of an NER method was followed by application of the soft-string matching technique to map newly discovered NEs to the most similar existing terms. This method discovered and mapped approximately 70000 new term variants. A further innovation of the BioLexicon is the inclusion of detailed information regarding the behaviour of verbs, which can leverage extraction of events.

In addition to including an extensive repository of biomedical terms and their arguments, the BioLexicon additionally incorporates detailed about typical syntactic and semantic patterns for domain-specific verbs, which is based on observable behaviour extracted from a corpus of biomedical texts [47].

The BioLexicon contains syntactic information (subcategorization frames) for 658 verbs, which were manually selected based on their particular relevance within the biomedical domain. For each verb, grammatical argument patterns (including modifier phrases) were extracted, based on the application of the Enju parser to a domain-specific corpus consisting of both biomedical abstracts on the subject of E. coli and full papers, totalling approximately 6 million words. Although modifier phrases (prepositional phrases and adverbials) are important in biomedical texts, they should only be considered to be arguments of the verb if there is sufficient evidence for this. Thus, a filter was used to ensure that rarely used patterns were not included in the lexicon. As each verb can occur with multiple patterns of syntactic arguments, a total of 1760 syntactic frames were extracted.

Semantic information about verbs was acquired based on a corpus of 677 abstracts that were manually annotated with events by domain experts, using a scheme almost identical to the one used for GREC, with the same 13 types of semantic roles [42]. The only difference is that whilst GREC is annotated with event instances, this second corpus was annotated with the specific purpose of extracting event frames to include within the BioLexicon.

Each extracted event frame was centred on a particular verb or nominalised verb. The subset of these frames that were centred on verbs for which grammatical information had been acquired was selected. This subset consisted of a total of 856 frames, centred on 168 verbs. A manual process was then used to link each argument in the syntactic subcategorisation frame its corresponding argument in the semantic frames. This resulted in 668 linked frames.

4. EVALUATION OF EVENT EXTRACTION

4.1 BioNLP'09 Shared Task

The development of event extraction systems that can reliably extract complex events involving multiple participants is an open research topic. However, the importance placed upon the development of such systems, and the desire of the community to push forward in this area have been demonstrated through the BioNLP'09 shared task [18]. Shared tasks involve teams from the community competing to analyze the same data within a common evaluation framework. They provide standard development and evaluation benchmarks, focusing the attention of the research community on timely issues and acting as a driver for the specification of new tasks and challenges. The BioNLP'09 shared task was the first to focus specifically on event extraction, which was based on protein biology event types.

The shared task evaluated the performance of systems not only in extracting primary event participants (i.e. THEME and CAUSE) but also secondary participants, including the source

and destination of the event. The results of the shared tasked showed that, although simple events can be extracted quite reliably using state of the art methods, more complex events involving multiple participants can currently only be extracted with less than 50% accuracy.

4.2 Evaluating the BioLexicon for Event Extraction

The BioLexicon has been evaluated within a challenging context, namely that of full parsing as part of the UKPubMedCentral (UKPMC) text mining services (http://ukpmc.ac.uk/), to locate and extract facts related to the biology domain. In practice, there are three components in the fact extraction process. Firstly, syntactic arguments of verbs in the texts are located through the application of the Enju parser to the texts. Only those verbs that are included in the BioLexicon are considered as potential textual "anchors" of events. These candidate events are further narrowed down by selecting only those in which an NE relevant to the domain appears in one of the arguments associated with the verb. As a final test, the syntactic argument pattern of the verb should be as predicted in the BioLexicon.

Whilst the primary use of the BioLexicon information in this context is as a filter, it also has a boosting effect on the range of facts to be considered. This is because modifier phrases (e.g., those which begin with prepositions) are explored, which would not be considered without its input. Where these modifier phrases contain recognised named entities, this can provide enough evidence for the extraction of a fact that would not otherwise be recorded. Consider the following example:

The pXPC3 plasmid codes for an XPC cDNA that is truncated by 160 bp from the N terminus compared with the wild-type XPC cDNA

Although the Enju parse result treats *code* as an intransitive verb (i.e. without a grammatical object), the information present in the BioLexicon allows the THEME role to be assigned to the prepositional phrase beginning with *for*.

The method described above has been evaluated through application to a test set of approximately 80,000 documents. Within these documents, only 62.7% of the instances of the verbs match verbal entries in the BioLexicon, thus illustrating its initial filtering effect. A still stronger filter is the requirement that a domain relevant NE should be present in one of the arguments. Applying this constraint results in only 16.9% of the total number of verb instances present in the text collection being extracted as facts. The experimental results also demonstrate, at least to some extent, the boosting effect achieved by using the verbal information in the BioLexicon, in that 9.7% of verb arguments are detected in prepositional modifier phrases, rather than in the arguments initially predicted by the parser output. These preliminary results provide compelling evidence that the BioLexicon can assist in building powerful tools for fact extraction within the biomedical domain.

5. EVENT INTERPRETATION

Although a large amount of work has been carried out on building resources and tools to facilitate extraction of events from biomedical texts, less attention has been paid to the way in which the extracted events should be interpreted. In addition to the event participants themselves, there is frequently additional information (or *meta-knowledge*) present within the context of the event that is vital to its correct interpretation. Examples of meta-knowledge include the type of evidence behind the event (e,g., does it represent a hypothesis, a well-established fact, etc.), whether there is any speculation expressed about the event, whether it is negated, etc.

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

5.1 Expression of Meta-Knowledge

To make the idea of meta-knowledge more concrete, consider Figure 1, which shows a set of eight simple sentences. 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 2 shows the typical structured representation of these events.

(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*

Figure 1 – Simple Sentences

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

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

Figure 2 – Structured Representation of E1 and E2

The event trigger words are underlined in each of the examples. The *expression* event (E1) is always indicated by the nominalised verb *expression*. However, the *positive regulation* event (E2) is expressed in a number of different ways, namely using the verbs *activate*, *increase* and *affect*, or the nominalised verb *effect.* 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.

The emboldened words and phrases in the example sentences help to show that the way in which the events should be interpreted can vary considerably. Most of the emboldened words affect the interpretation of the event E2, which is the main event in the sentence. However, in (S7) the interpretation of E1 is altered.

Sentences (S1), (S5), (S7), and (S8) all describe experimental observations. In most of these, the presence of a particular word (i.e., *found*, *shown* and *observed*) marks the E2 positive regulation as being an observation. In (S5), however, it the use of the past tense on the word on which the positive regulation event is centred (i.e., *increased*) that marks it as an observation.

Although all 4 events mentioned above represent observations, each of their interpretations is still slightly different. The different between (S1) and (S8) is the source of the information. The presence of the word *we* in (S1) indicates an observation as part of the current study, whilst in (S8), *previous studies* denotes an observation originally reported outside of the current paper. Thus, in (S1), the positive regulation can be considered as "new" knowledge, but (S8), the knowledge reported is "old". Whether such a difference is important will depend on the task being undertaken by the user. For example, database curators looking only for new knowledge might only be interested in (S1).

In (S5) and (S7) the difference in interpretations concerns event intensity, through the words *slightly* (i.e., low intensity) and *significant* (i.e., high intensity), respectively. The recognition of such information about events may be important, for example, when performing a comparison of different experimental methods. In (S5), the intensity applies to E2, whilst in (S7), the intensity applies to E1, as this is the only event that appears in the sentence.

The positive regulation event is (S4) can be taken as a well-established fact within the field, according to the presence of the word *known*. In a system that is looking for contradictions, events that contradict this well-established fact are potentially more serious than, say, a contradiction of new experimental outcomes (e.g., (S1)), which could later be disputed by other experts within the field.

All the events described above can be seen as reporting factual information. In this respect, (S2) is quite different. The presence of the word *examined* serves to indicate that the positive regulation event is under examination, and so it is not known whether or not it is true.

Sentences (S3) and (S6) should also not be considered as facts. Rather, the presence of the word *suggests* denotes that E2 is being stated as a somewhat tentative analysis of results on the part of the author. In (S6), the author uses the word *might* to increase the amount of speculation about the truth of the event. In (S3), the conclusion is different: the author concludes is that the positive regulation event is unlikely to happen, indicated by the use of the word *no*. Hence, this is a negative event.

From the above sentences, it is possible to isolate at least five important pieces of contextual information which can be regularly identified about events, which somehow modify their default interpretation:

1) What kind of evidence is there for the event, e.g. has it been experimentally observed, inferred from experimental results, is a well established fact, or is it a hypothesis whose truth has yet to be determined?
2) How certain is the author about whether the event is true?
3) Is the event positive, or is it negated (through the use of *no, not* etc.)
4) What is the intensity or magnitude of the event?
5) What is the source of the information contained within the event? Is it reported in the current paper or another paper?

5.2 Meta-Knowledge Annotation of Bio-Events

Existing event annotated corpora within the biomedical domain contain few annotations that relate to their interpretation. Negations are annotated in BioInfer and GENIA. Three different levels of certainty are also annotated for GENIA events. However, negation and speculation clue words are not annotated in these corpora. Negation and speculation were also addressed in one of the subtasks of the BioNLP'09 shared task, but in a fairly basic way. The only requirement was to recognise whether events were negated and/or contained expressions of speculation, without having to identify, e.g. the level of speculation. Only 6 out of the 24 participating teams attempted this task and the highest accuracy was around 25%. This was attributed to the lack of annotated clue phases in the training corpus [18].

More extensive interpretation-focussed annotation has been carried out within the domain at either the sentence level (e.g., [48]) or sentence-fragment level (e.g., [50]). However, these annotations cannot be used straightforwardly to assign interpretations to bio-events. Often, a sentence will contain several bio-events (e.g. both an experimental method *and* the results of applying this method), each of which has a different interpretation. If an expression of speculation is present (e.g. the word *might*), this may affect only certain events in a sentence.

Based on the above, we have designed a multi-dimensional annotation scheme to capture various aspects of meta-knowledge expressed for bio-events [28]. Our 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, which may be important according to the task being undertaken by the biologist.

The annotation task consists of assigning an appropriate value for each dimension, as well as marking the textual evidence for this assignment. This latter part of the task is important to train systems to perform meta-knowledge identification successfully, given the difficulties faced in the negation/speculation part of the BioNLP'09 shared task, where such annotations were not present in the training data.

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. This aspect of our scheme is further discussed in section 5.2.1.

Figure 3 provides an overview of the annotation scheme. The boxes with the light-coloured background correspond to

information that is common to most bio-event annotation schemes, whilst the boxes with the darker backgrounds correspond to our proposed meta-knowledge annotation dimensions and their possible values. Below, we provide brief details of each annotation dimension.

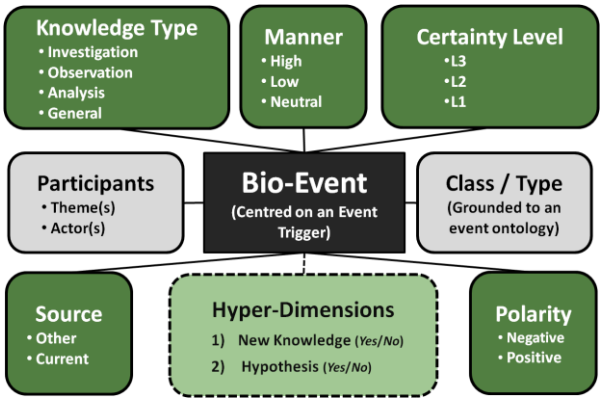

Figure 3 - Bio-Event Annotation

Knowledge Type (KT): Captures the general information content of the event. Each event is classified as either: *Investigation* (enquiries and examinations etc.), *Observation* (direct experimental observations), *Analysis* (inferences, interpretations and conjectures etc.) or *General* (facts, processes, states or methodology)

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* (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.

Manner: Captures information about the rate, level, strength or intensity of the event, using three values: *High* (increase in rate/intensity), *Low* (decrease in rate/intensity) or *Neutral* (no indication of rate/intensity).

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

5.2.1 Hyper-Dimensions

A defining feature of our annotation scheme is that additional information (hyper-dimensions) can be inferred by considering combinations of some of the explicitly annotated dimensions. These are as follows:

New Knowledge: A combination of the values of *Source, KT* and *CL* dimensions can be used to isolate those events representing new knowledge. Specifically, new knowledge corresponds to events with a *KT* value of *Observation* or *Analysis* carried out as part of the current study (i.e., *Source=Current*). If *KT=Analysis*, then the event should only be classed as new knowledge if it represents a straightforward interpretation of results (i.e. *CL=L3*), rather than something more speculative.

Hypothesis: Events that represent hypotheses can be isolated by considering *KT* and *CL* values. Events with a *KT* value of *Investigation* can always be assumed to be a hypothesis. However, if the *KT* value is *Analysis,* then only those events with a *CL* value of L1 or L2 should be considered as hypotheses.

5.3 Feasibility and Application

An initial evaluation of the annotation scheme has been performed through the annotation of 70 abstracts randomly chosen from the GENIA Pathway Corpus, containing a total of 2,603 annotated bio-events. Two annotators performed the annotation using a comprehensive set of annotation guidelines developed following a detailed analysis of the various bio-event corpora and the output of an initial case study [28].

The evaluation results have shown high inter-annotator agreement and a sufficient number of annotations along each category in every dimension. The favourable results of this experiment have confirmed the feasibility and soundness of the annotation scheme, and have paved the way for a large scale annotation effort involving multiple independent (i.e. non-author) annotators.

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

6. Conclusion

In this paper, we have described how text mining can help biologists to search and locate relevant information within the literature in a much more effective and efficient manner than is possible using a traditional search engine that performs keyword searches over unstructured documents.

Text mining techniques can be applied to biomedical texts to extract structured, semantically-oriented event representations of the biomedical knowledge contained within the texts. Queries can then be applied to these extracted events, rather than on the unstructured documents. Such queries can themselves be structured, allowing specifications of exactly which search terms should be related to each other, and how.

Extraction of events is a complex process requiring a number of text mining technologies, including NER and deep parsing. NER is important to ensure that only events containing biologically relevant entities are recognised, whilst parsing helps to identify potential event participants through syntactic relations. Annotated corpora of events are important for training systems to recognise events and their participants, as they provide direct evidence of how events manifest themselves in text. Computational lexicons such as the BioLexicon can further enhance performance, in providing detailed information about the idiosyncratic behaviour of verbs on which events are often centred.

Information regarding the intended interpretation of events is also important. Our proposed meta-knowledge annotation scheme for events and ongoing work to produce a large corpus of events annotated according to this scheme will form an important first step in allowing systems to be trained to recognise interpretative information about events from huge repositories.

7. REFERENCES

[1] Ananiadou, S., Friedman, C. and Tsujii, J. (eds.) Special Issue on Named Entity Recognition in Biomedicine. *Journal of Biomedical Informatics*, 37, 6 (2004).

[2] Ananiadou, S., Kell, D. B. and Tsujii, J. Text mining and its potential applications in systems biology. *Trends Biotechnol*, 24, 12 (Dec 2006), 571-579.

[3] Ananiadou, S. and Nenadic, G. *Automatic Terminology Management in Biomedicine.* Artech House Books, 2006.

[4] Ananiadou, S., Pyysalo, S., Tsujii, J. and Kell, D. B. Event extraction for systems biology by text mining the literature. *Trends Biotechnol.*, 28, 7 (Jul 2010), 381-390.

[5] 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. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25 (2000), 25-29.

[6] Bjorne, J., Ginter, F., Pyysalo, S., Tsujii, J. and Salakoski, T. Complex event extraction at PubMed scale. *Bioinformatics*, 26, 12 (Jun 2010), i382-390.

[7] Bodenreider, O., Burgun, A. and Rindflesh, T. Assessing the Consistency of a Biomedical Terminology through Lexical Knowledge. *International Journal of Medical Informatics*, 67, 1-3 (2002), 85-95.

[8] Browne, A. C., Divita, G., Aronson, A. R. and McCray, A. T. UMLS language and vocabulary tools. In *Proceedings of the Proceedings of the AMIA Annual Symposium* (Washington DC, USA, 2003).

[9] Dolbey, A., Ellsworth, M. and Scheffczyk, J. BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies. In *Proceedings of the Proceedings of KR-MED 2006: Biomedical Ontology in Action* (Baltimore, USA, 2006).

[10] Eaton, A. D. HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res*, 34, Web Server issue (Jul 2006), W745-747.

[11] Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. and Pappas, G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *Faseb J*, 22, 2 (Feb 2008), 338-342.

[12] Hearst, M. A. *Search User Interfaces*. Cambridge University Press, 2009.

[13] Hearst, M. A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M. A. and Ye, J. BioText Search Engine: beyond abstract search. *Bioinformatics*, 23, 16 (Aug 2007), 2196-2197.

[14] Hirschman, L. and Blaschke, C. *Evaluation of Text Mining in Biology*. Vol 9, Artech House Books, 2006.

[15] Hirschman, L., Colosimo, M., Morgan, A. and Yeh, A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, 6 Suppl 1 (2005).

[16] Hoffmann, R. and Valencia, A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21 Suppl 2 (Sep 2005), ii252-258.

[17] Hull, D., Pettifer, S. R. and Kell, D. B. Defrosting the digital library: bibliographic tools for the next generation web. *PLoS Comput Biol*, 4, 10 (Oct 2008), e1000204.

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

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

[20] Kipper-Schuler, K. *VerbNet: a broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

[21] Krauthammer, M. and Nenadic, G. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics, Special Issue on Named Entity Recognition in* Biomedicine (2004).

[22] Liu, H., Hu, Z. Z., Zhang, J. and Wu, C. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22, 1 (Jan 2006), 103-105.

[23] Miwa, M., Saetre, R., Kim, J. D. and Tsujii, J. Event extraction with complex event classification using rich features. *J. Bioinform. Comput. Biol.*, 8, 1 (Feb 2010), 131-146.

[24] Miyao, Y., Ninomiya, T. and Tsujii, J. Corpus-oriented grammar development for acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In *Proceedings of the Proceedings of IJCNLP 2004* (2004).

[25] Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T. and Tsujii, J. Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. *Annual Meeting- Association for Computational Linguistics*, 2 (2006), 1017-1024.

[26] Miyao, Y. and Tsujii, J. Feature Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics, MIT Press*, 34(1) (2008), 35-80.

[27] Muin, M. and Fontelo, P. Technical development of PubMed interact: an improved interface for MEDLINE/PubMed searches. *BMC Med Inform Decis Mak*, 6, 36 (2006).

[28] Nawaz, R., Thompson, P., McNaught, J. and Ananiadou, S. Meta-Knowledge Annotation of Bio-Events. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)* (Malta, 17-23 May, 2010).

[29] Nobata, C., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J. and Ananiadou, S. Kleio: a knowledge-enriched information retrieval system for biology. In *Proceedings of the 31st Annual International ACM SIGIR Conference*, (2008), 787-78.

[30] Okazaki, N., Ananiadou, S. and Tsujii, J. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26, 9 (May 2010), 1246-1253.

[31] Okazaki, N., Ananiadou, S. and Tsujii, J. i. Building a High Quality Sense Inventory for Improved Abbreviation Disambiguation. *Bioinformatics* 26, 9 (2010).

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

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

[34] Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. and Jimeno, A. Text processing through Web services: calling Whatizit. *Bioinformatics*, 24, 2 (Jan 2008), 296-298.

[35] Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P. EBIMed--text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23, 2 (Jan 2007), e237-244.

[36] Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, C. and Scheffczyk, J. *FrameNet II: Extended Theory and Practice*. 2006.

[37] Sasaki, Y., Montemagni, S., Pezik, P., Rebholz-Schuhmann, D., McNaught, J. and Ananiadou, S. BioLexicon: A Lexical Resource for the Biology Domain. In *Proceedings of the Proceedings of the Third International Symposium on Semantic Mining in Biomedine (SMBM 2008)* (Turku, Finland, 2008).

[38] Sasaki, Y., Tsuruoka, Y., McNaught, J. and Ananiadou, S. How to make the most of named entity dictionaries in statistical NER *BMC Bioinformatics*, 9 Suppl 11 (2008).

[39] Sasaki, Y., Tsuruoka, Y., McNaught, J. and Ananiadou, S. How to make the most of NE dictionaries in statistical NER. *BMC bioinformatics*, 9 Suppl 11 (2008).

[40] States, D. J., Ade, A. S., Wright, Z. C., Bookvich, A. V. and Athey, B. D. MiSearch adaptive pubMed search tool. *Bioinformatics*, 25, 7 (Apr 2009), 974-976.

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

[42] Thompson, P., McNaught, J., Ananiadou, S., Montemagni, S., Trabucco, A. and Venturi, G. Building a bio-event annotated corpus for the acquisition of semantic frames from biomedical corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)* 2008.

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

[44] Tsuruoka, Y., McNaught, J. and Ananiadou, S. Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics*, 9 Suppl 3 (2008).

[45] Tsuruoka, Y., McNaught, J., Tsujii, J. and Ananiadou, S. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23, 20 (Oct 2007), 2768-2774.

[46] UniProt Consortium The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, 38, Database issue (2010), D142-D148.

[47] Venturi, G., Montemagni, S., Marchi, S., Sasaki, Y., Thompson, P., Mcnaught, J. and Ananiadou, S. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In *Proceedings of the Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing* (Mexico City, Mexico, 2009). Springer-Verlag.

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

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

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

[51] Wren, J. D., Chang, J. T., Pustejovsky, J., Adar, E., Garner, H. R. and Altman, R. B. Biomedical term mapping databases. *Nucleic Acids Research*, 3, Database Issue: D289-293 (2005).

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