A Specialised Verb Lexicon as the Basis of Fact Extraction in the Biomedical Domain

C.J. Rupp, Paul Thompson, William Black, John McNaught and Sophia Ananiadou
National Centre for Text Mining, University of Manchester, UK
c.j.rupp@cs.man.ac.uk,
{paul.thompson,william.black,john.mcnaught,sophia.ananiadou}@manchester.ac.uk

Abstract

The BioLexicon is a standardised, reusable, lexical and conceptual resource suitable for advanced biomedical text mining. One of the unique features of the BioLexicon is the incorporation of rich syntactic and semantic patterns for a wide range of domain-relevant verbs, which have been acquired semi-automatically from biomedical corpora. Such types of information can be highly beneficial for information and fact extraction applications. In this paper, we describe the collection of the verb-specific information for inclusion in the BioLexicon, and explain how it is being employed in a specific scenario (the UKPMC project) to leverage fact-based information extraction on a large collection of biomedical papers.

1 Introduction

Information Extraction (IE) applications that focus on extraction of event information require sophisticated lexical resources, which include both syntactic patterns for verbs, and their corresponding semantic interpretations (in terms of semantic roles). The BioLexicon (Sasaki et al., 2008) is the first large-scale, specialised lexical resource that includes such information for a wide range of domain-relevant verbs.

The BioLexicon is being evaluated within the context of the UK PubMed Central (UKPMC) project, where it is used as a fulcrum to leverage fact-based information extraction over a large collection of approximately 1.8M research articles in the biomedical domain.

The UKPMC project is developing a specific search portal for UK researchers, which accesses the PubMed Central repository of biomedical research, and adds additional functionalities, including text mining capabilities.

Among the text mining applications under development is a semantic search engine that allows specific facts to be located within the document collection. To support this application, an extensive index of analysed facts occurring within the collection is under compilation. The detailed information provided in the BioLexicon regarding the behaviour of verbs in the biomedical domain forms the primary criterion for recognising facts and extracting their constituents.

In the first part of this paper, we describe the motivation for the construction of the BioLexicon, followed by a description of the collection of the verb-specific information contained within it. In the second part, we explain in more detail the method by which the BioLexicon is employed in the context of the UKMPC project, and provide an initial evaluation of this method.

2 Lexical Resources for IE

A number of large-scale computational lexicons containing syntactic and semantic information for verbs and other parts-of-speech have been developed for general English language, e.g., FrameNet (Rupenhoffer et al., 2006), Propbank (Palmer et al., 2005) and VerbNet (Kipper-Schuler, 2005). However, these resources are not well suited for use in IE systems that operate in specialized domains such as biomedicine, and may lead to incorrect analyses.

Descriptions of events in biomedical texts have a number of domain-specific features. Firstly, there are verbs that appear rarely in general language texts (e.g., *phosphorylate*), and hence are not accounted for in the general language resources. Secondly, verbs that occur frequently in both general language and biomedical texts often have different syntactic or semantic properties in each domain, e.g., differing numbers of arguments (Wattarujeekrit et al., 2004) or different meanings. In addition, strongly selected modifiers (such as location, manner and timing), are considered to be much more important to the correct interpretation of biomedical events than general language events (Tsai et al., 2007). This is exemplified in the following sentence, which specifies both a manner and a location for the event described by the verb *directs*:

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

Although domain-specific extensions of general language computational lexicons have been attempted, e.g., BioFrameNet (Dolbey et al., 2006) and PASBio (Wattarujeekrit et al., 2004), their coverage is very limited. The SPECIALIST lexicon (Browne et al., 2003) is a larger resource containing biomedical vocabulary. Although syntactic complementation patterns are included for verbs, they are somewhat limited, and based on general language patterns. In addition, no semantic information for verbs is provided.

3 The BioLexicon

The BioLexicon[1] (Sasaki et al., 2008) is a standardised, reusable, lexical and conceptual resource suitable for advanced biomedical text mining, containing over 2.2M lexical entries, with a particular emphasis on gene regulation.

Whilst the vast majority of entries in the BioLexicon correspond to biomedical terms, a major design criterion was to include syntactic and semantic patterns for a wide range of domain-relevant verbs, in order to address the previous lack of a suitable resource. The BioLexicon thus incorporates 658 domain-relevant verbs, all of which are accompanied by syntactic subcategorization frames, and 168 of which include semantic event frames, as well as explicit linking between the syntactic and semantic levels.

A corpus-based approach was taken to the construction of the verbal part of the lexicon to ensure that the behaviour of the verbs recorded in the lexicon reflects the way they are used in domain-specific texts. In contrast to the manual construction of many other lexical semantic resources, the verbal information in the BioLexicon was derived semi-automatically, using different techniques and different sizes of corpora to obtain each type of information.

3.1 Syntactic Subcategorisation Frames

The extraction of subcategorisation frames (SCFs) (Venturi et al., 2009) used an unsupervised learning technique, applied to a corpus of approximately 6M tokens (both MEDLINE abstracts and full biomedical papers) on the subject of E. coli. The corpus was automatically annotated for predicate-argument structure using a version of the Enju parser tuned to biomedical texts (Hara et al, 2005). Based on the parse results, observed dependency sets (ODSs) were computed for each verbal occurrence and used as the basis of the SCFs.

Each ODS is represented as a set of dependencies described in terms of relation type (e.g. ARG1, ARG2, etc.). The order of the dependencies in each ODS is normalised and does not reflect their order of occurrence in context. According to their importance in biomedical events, the induced SCFs include strongly-selected modifiers as well as strongly selected arguments in the description of biomedical events.

For each ODS, the conditional probability given the verb was computed. Thresholding based on this probability was used to filter out noisy frames (i.e., frames containing not only arguments and strongly selected modifiers, but also adjuncts) as well as possible errors of either parsing or ODS extraction. The remaining 1760 ODSs (distributed amongst the 658 verbs) were selected as SCFs for inclusion in the BioLexicon.

3.2 Semantic Event Frames

The extraction of event frames was carried out on a subset (677 abstracts) of the corpus used for SCF extraction. Each abstract was manually annotated with gene regulation events, centred on both verbs and nominalised verbs, by a group of domain experts (Thompson et al., 2008). For each event, semantic arguments occuring within the same sentence were labelled with both semantic roles and named entity (NE) types.

Although somewhat comparable to the GENIA event annotation (Kim et al., 2008), our annotation differs in that it was geared specifically towards the acquisition of semantic frame information for verbs, using a richer set of semantic roles to capture detailed information regarding verb behaviour. Our corpus uses a total of 13 roles (compared to 6 in GENIA), which are intended to characterise all the sublanguage semantic arguments of relevant events.

The semantic roles used in the BioLexicon are event-independent, and constitute a closed set, which is advantageous in facilitating generalization over different types of events (Cohen and Hunter 2006; Merlo and Plas 2009). Although application of a closed semantic role set to general language events may be problematic (Palmer et al., 2005), the use of such a set is more viable in a restricted domain, as domain-specific definitions can be provided for each semantic role type.

[1] http://catalog.elra.info/product_info.php?products_id=1113

Role Name	Description	**Example** ([…] = semantic argument, small capitals = 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 radition [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 E. Coli.
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 is activated by BINDING of the cyclic AMP (cAMP)-cAMP receptor protein (CRP) complex to [a CRP binding site]
TEMPORAL	Situates event in time/ w.r.t. another event	The Alp protease activity is DETECTED in cells [after introduction] of plasmids
CONDITION	Environmental conditions/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	HyfR ACTS as [a formate-dependent regulator]
DESCRIPTIVE-THEME	Descriptive information about THEME of event	The FNR protein RESEMBLES [CRP].
PURPOSE	Purpose/reason for the event occurring	The fusion strains were USED [to study] the regulation of the cysB gene

Table 1: Semantic roles and definitions

Our semantic roles are based largely on the verb-independent roles used in VerbNet (Kipper-Schuler, 2005) and SIMPLE (Lenci et al, 2000). Through the examination of a large number of relevant events within MEDLINE abstracts, in consultation with biologists, it was concluded that arguments of gene regulation events may be characterised using a subset of these general language roles, with some name changes to make them more easily understandable to biologists, and with the addition of the domain-specific CONDITION role, corresponding to descriptions of environmental conditions. The full set of roles is shown in Table 1.

NE categories are organised into 5 different hierarchies, corresponding to the following 5 supercategories: *DNA, PROTEIN, EXPERIMENTAL, ORGANISMS* and *PROCESSES*. The categories are mapped to classes in the Gene Regulation Ontology (GRO) (Beisswanger et al, 2008).

A set of 856 verb-specific semantic frames was extracted from the annotated corpus for inclusion in the BioLexicon. We have chosen to create verb-specific frames, as these allow more detailed argument specifications than those resources that group verbs into classes (e.g.,. VerbNet, FrameNet). The importance within the domain of phrases that identify location, manner, timing and condition mean that individual verbs can behave idiosyncratically.

Extracted semantic frames include the semantic roles annotated, in addition to NE types, if available. These allow selectional restrictions to be applied to the fillers of each role. An example event frame is as follows:

```
activate(Agent=>Protein,
         Theme=>DNA)
```

3.3 Linking Syntactic and Semantic Frames

Syntactic arguments of predicates have been manually linked to their semantic counterparts in the event frames, in order to facilitate the automatic labelling of syntactic arguments of verbs with semantic roles. This step was carried out for the 168 verbs for which both subcategorisation and event frame information was available, taking into account the following types of information:

a) General linguistic constraints regarding the alignment of hierarchies of semantic roles and grammatical functions. Given a semantic role hierarchy (agent>theme ...) and a grammatical functions hierarchy (subject>object ...), the mapping usually proceeds from left to right;

b) A list of 'prototypic' grammatical realisations of semantic arguments;

c) General language repositories of individual semantic frames containing both syntactic and semantic information.

4 Fact-Based Information Extraction

The sheer volume of publications in biomedicine has made it a focus for text mining research. Much of this activity involves named entity recognition (NER), i.e., the identification of technical terms and designations relevant for the domain. Text mining systems may either manage the documents themselves, i.e. information retrieval, or the information contained within the documents, i.e. information extraction (IE). IE applications aim to locate relations between entities, e.g., Hoffman and Valencia (2004). These relations may be evidenced by proximity in the text, or inferred based on domain knowledge. The most specific relations are the claims explicitly made in the text detailing the research itself.

We aim to support searching over the evidence and claims presented in research papers by indexing the relations that occur at the lexical level. We refer to the combinations of lexical relations and arguments, which typically centre on verbs or deverbal nominalizations, as *facts*[2].

The detailed information encoded about verbs in the BioLexicon forms the keystone of an IE method applied to the UKPMC corpus. In order to support queries against this collection focussing on specific evidence presented in the text, we analyse the verbal relations, along with their argument structure and predicted modifiers. We extract representations of the key facts, and then index these for efficient query and retrieval. The BioLexicon provides the information that ultimately decides which constructions are recorded as facts.

There are three knowledge sources used in the fact extraction process:

The papers are syntactically analysed using the Enju parser, which is the same parser used in the development of the BioLexicon. It has been optimised for the biomedical domain by the use of a parse preference model (Hara et al., 2005), meaning that we can be confident in selecting only the highest rated parse for each sentence. The size of the collection would make the considering competing parses impractical.

The extended verb frames of the BioLexicon, which provide patterns of argument structure and systematic modification, are used as predictions of the arguments and modifier structures that can identify relevant facts within the domain.

Within the UKPMC project, a suite of standardised NER recognisers for various classes of named entities is used across all applications. The NER results play a significant role in determining which facts to extract and index.

The fact extraction process consists of three steps, each refining the set of potential facts more precisely. Firstly, we locate within the Enju parse result those verbs with corresponding entries in the BioLexicon. Next, we require that at least one of the named entities recognised by the NER components be involved in the relation centred on the verb, either as an argument or as a predicted modifier. Finally, we ensure that whole construction is consistent with a verb frame definition in the BioLexicon.

As an example, consider the two syntactic frames provided in the BioLexicon for *induce*:

```
1) induce,ARG1#ARG2#
2) induce,ARG1#ARG2#PP-in#
```

NER recognises *mutant p53* as a protein. This allows us to add to our index sentences where this protein appears with verb *induce*, either as the subject, object or a prepositional modifier headed by *in*, e.g.:

This scenario suggests that mutant p53 could use different mechanisms to ***induce*** *malignant properties in epidermal keratinocytes.*

The overall conclusion from our work is a direct relationship between chemoresistance ***induced*** *by mutant p53 and its transactivation ability.*

[2] However, we are aware that not all of the claims they express are factual

The primary search domain of this extraction process is the analysis tree provided by the Enju parser. The alignment of recognised named entities with analysed constituents is performed via the standoff annotations provided by each component. The relevant sections of text are retrieved and recorded, in order to present search results from the index. The information from the BioLexicon is the primary filter and determines the final choice of facts to be indexed, but the results of both parsing and NER make a significant contribution.

5 Evaluation of the BioLexicon for Fact Extraction

An initial quantitative evaluation of the method described above has been carried out on a subset of the UKPMC corpus, consisting of approximately 80,000 documents.

On the one hand, the BioLexicon is a strong filter, in that only the verbs it recognises are accepted as the basis of a fact. This is restrictive in that only a certain part of the domain covered by the collection is within the remit of the BioLexicon, i.e., gene regulation. The evaluation results confirm this filtering effect: only 62.7% of the instances of the verbs present in the document collection matched verbal entries in the BioLexicon. A still stronger filter is the requirement that a domain relevant NE should be present in one of the arguments, resulting in only 16.9% of the total verb instances present in the text collection being extracted as facts.

On the other hand, the lexicon also has a boosting effect on the fact base, since modifier phrases 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 *codes* as an intransitive verb, the information in the BioLexicon allows the THEME role to be assigned to the PP headed by *for.*

This boosting effect is demonstrated in the evaluation results: 9.7% of verb arguments are detected in prepositional modifier phrases, rather than in the arguments initially predicted by the parser output.

In addition to the argument and modification patterns predicted in the BioLexicon, the fact index also records patterns of negation and some other scoped modifications that are independent of the lexical predictions. We are thus able to distinguish between logically related facts retrieved in a query-based application.

5.1 Conclusion and Further Work

This paper has described the verbal component of the BioLexicon, which is a unique resource comprising rich linguistic information suitable for text mining applications operating within the biomedical domain. The corpus-driven nature of the acquisition of both syntactic and semantic information for verbs aims to facilitate the accurate identification of events, together with their participants and the semantic roles assigned to them. Such comprehensive information is not currently available in any comparable domain-specific resource.

The BioLexicon is at the heart of an IE method that is being employed to facilitate fact-based querying over a large collection of biomedical documents as part of the UKPMC project. The preliminary results provide compelling evidence that the BioLexicon can assist in building powerful tools for fact extraction within the biomedical domain.

We are currently in the process of developing applications based on the fact index extracted for the UKPMC corpus (see Black et al. (2010)).

The utility of the BioLexicon has also been shown to extend beyond IE applications; the recognition of multiword terms in the lexicon can help with a number of NLP tasks in the biomedical domain including POS tagging and syntactic parsing (Sasaki et al., 2009) and improving the performance of information retrieval (Sasaki et al., 2010).

Acknowledgments

This research was supported by the European Commission IST project FP6-028099 (BOOTStrep) and the UKPubMedCentral (UKPMC) project. UK PubMed Central is funded by: Arthritis Research Campaign, BBSRC, British Heart Foundation, Cancer Research UK, Chief Scientist Office, Department of Health - National Institute of Health Research, Medical Research Council, Wellcome Trust. We also thank ILC-CNR, Italy for their work on the production of the subcategorisation frames.

References

E. Beisswanger, V. Lee, J.J Kim, D. Rebholz-Schuhmann, A. Splendiani, O. Dameron, S. Schulz and U. Hahn. 2008. Gene Regulation Ontology (GRO): design principles and use cases. *Studies in Health Technology and Informatics*, 136:9-14.

W.J. Black, C.J. Rupp, C. Nobata, J. McNaught, J. Tsujii and S. Ananiadou. 2010. High-Precision Semantic Search by Generating and Testing Questions. In *Proceedings of the UK e-Science All Hands Meeting 2010.*

A.C. Browne, G. Divita, A.R. Aronson and A.T. McCray. 2003. UMSL Language and Vocabulary Tools: AMIA 2003 Open Source Expo. In *Proceedings of AMIA Annual Symposium 2003*, page 798

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

A. Dolbey, M. Ellsworth and J. Scheffczykx. 2006. BioFrameNet: A Domain-Specific FrameNet Extension with Links to Biomedical Ontologies. In *Proceedings of KR-MED,* pages 87-94.

T. Hara, Y. Miyao and J. Tsujii. 2005. Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In *Proceedings of IJCNLP*, pages 199-210

R. Hoffmann, R. and A. Valencia. 2004. A Gene Network for Navigating the Literature. *Nature Genetics* 36:664

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

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

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

P. Merlo and L. Plas. 2009. Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? In *Proceedings of ACL-IJCNLP 2009*, pages 288-296

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

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

Y. Sasaki, S. Montemagni, P. Pezik, D. Rebholz-Schuhmann, J. McNaught and S. Ananiadou 2008. BioLexicon: A Lexical Resource for the Biology Domain. In *Proceedings of the SMBM 2008*, pages 109-116.

Y. Sasaki, P. Thompson, J. McNaught and S. Ananiadou. 2009. Three BioNLP Tools Powered by the BioLexicon. In *Proceedings of EACL Demonstration Session*, pages 61-64

Y. Sasaki, J. McNaught and S. Ananiadou. 2010. The value of an in-domain lexicon in genomics qa. *Journal of bioinformatics and computational biology*, 8(1):147-161.

P. Thompson, P. Cotter, S. Ananiadou, J. McNaught, S. Montemagni, A. Trabucco and G. Venturi .2008. Building a Bio-Event Annotated Corpus for the Acquisition of Semantic Frames from Biomedical Corpora. In *Proceedings of LREC 2008*, pages 2159-2166.

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

G. Venturi, S. Montemagni, S. Marchi, Y. Sasaki, P. Thompson, J. McNaught, and S. Ananiadou. 2009. Bootstrapping a Verb Lexicon for Biomedical Information Extraction. In *Proceedings of CICLing 2009*, pages 137-148.

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