BEL networks derived from qualitative translations of BioNLP Shared Task annotations

Juliane Fluck[1*], Alexander Klenner[1], Sumit Madan[1], Sam Ansari[2], Tamara Bobic[1,3], Julia Hoeng[2], Martin Hofmann-Apitius[1,3], Manuel C. Peitsch[2]

[1]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany.

[2]Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland.

[3]Bonn-Aachen International Centre for Information Technology, Dahlmannstr. 2, Bonn, Germany

{jfluck, smadan, aklenner, tbobic, mhofmann-apitius}@scai.fraunhofer.de,
{sam.ansari, julia.hoeng, manuel.peitsch}@pmi.com

Abstract

Interpreting the rapidly increasing amount of experimental data requires the availability and representation of biological knowledge in a computable form. The Biological expression language (BEL) encodes the data in form of causal relationships, which describe the association between biological events. BEL can successfully be applied to large data and support causal reasoning and hypothesis generation.

With the rapid growth of biomedical literature, automated methods are a crucial prerequisite for handling and encoding the available knowledge. The BioNLP shared tasks support the development of such tools and provide a linguistically motivated format for the annotation of relations. On the other hand, BEL statements and the corresponding evidence sentences might be a valuable resource for future BioNLP shared task training data generation.

In this paper, we briefly introduce BEL and investigate how far BioNLP-shared task annotations could be converted to BEL statements and in such a way directly support BEL statement generation. We present the first results of the automatic BEL statement generation and emphasize the need for more training data that captures the underlying biological meaning.

1 Introduction

Currently a lot of effort is made to extract information from scientific articles and encode the relevant parts in machine-readable language. In order to tackle these tasks, curators must be experts in both biological domain and computational representation of knowledge.

With the introduction of BEL, a new knowledge coding convention was made available, thus simplifying the curation process and ensuring machine readability[1]. BEL was initially designed and used in 2003 by Selventa (operating as Genstruct® Inc. at the time) to capture relationships between biological entities in scientific literature (Slater and Song 2012). It is flexible enough to store content from multiple knowledge layers and a broad range of analytical and decision-supporting applications. Knowledge bases encoded in BEL are suitable for querying, interpreting, reasoning and visualising of networks.

BEL represents scientific findings by capturing causal and correlative relationships in a given context, including information about the biological system and experimental conditions. The supporting evidences are captured and linked to the publication references. It is specifically designed to adopt external vocabularies and ontologies, and therefore represents life-science knowledge in language and schema known by the community. Entities in BEL statements are mapped to widely accepted namespaces, which specify a set of domain entities (e.g., HGNC[2], CHEBI[3]). Continuous development and commercial use in more than 80 life science projects in the last ten years qualify BEL as suitable for displaying causal networks for both humans and computers. Various networks built in BEL were mainly focusing on disease mechanisms (Schlage

[1] http://wiki.openbel.org/display/BLD/BEL+Language+Documentation+v1.0+-+Current

[2] http://www.genenames.org/

[3] http://www.ebi.ac.uk/chebi/

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 80–88,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

et al., 2011) and are used for causal reasoning (Chindelevitch et al., 2012, Huang et al., 2012 and Selventa 2012). Since 2012, BEL is also available in the public domain through the OpenBEL consortium. The OpenBel portal[4] defines the BEL language standard and provides formatted content and compatible tools for research.

The necessary information to develop a BEL knowledge base is currently harvested mainly by manual translation of literature into BEL statements. To support automated extraction of statements by text mining techniques, additional efforts and adaptations of existing text mining platforms are necessary.

The BioNLP community has developed various approaches, which may already support the automated extraction of BEL statements. To estimate how far current tools can generate BEL relationships, we focused on the BioNLP shared tasks series[5]. The BioNLP-shared tasks specify fine-grained information extraction tasks for biologically relevant targets, mainly centred on proteins and genes. In the two previous events, BioNLP-ST 2009 and 2011, more than 30 teams participated with their systems, a number of which are available as open source. In BioNLP-ST 2013 series, additional training data for pathway curation including chemical entities is available.

The organizers develop a linguistically based event representation and provide annotated training and test data to the participants. The annotated events in training data can directly be used for comparison with BEL definitions and available BEL statements. If the conversion of said event annotations to BEL statements (and vice versa) is successful on the semantic level, we have a promising opportunity to support both domains. Information encoded in the BEL statements in combination with corresponding evidence sentences could be used as training data to support further tool development.

2 Related Network Representations

For pathway representations there exist two widely adopted machine readable representations: Systems Biology Markup Language (SBML)[6] (Hucka et al., 2003) and Biological Pathway Exchange (BioPAX) (Demir et al., 2010). SBML is an XML-based data exchange format that supports a formal mathematical representation of chemical reactions including kinetic parameters. BioPAX is an RDF/OWL-based standard language enabling integration, exchange, visualization, and analysis of biological pathway data. Pathway representations in BioPax were already compared to the BioNLP-ST representations (Ohta et al., (1) 2011) and let to the introduction of the Pathway curation task in 2013[7]. For this task additional entity types and event types were proposed and resulted in a set of new annotations (Ohta et al., (2) 2011). A comparison between BEL and BioPax can be found at the OpenBEL Portal[8]. BioPAX focuses on pathway construction and partly may require more information than available in most publications. BEL's design enables the representation of causal relationships across a wide range of mechanistic detail and between the levels of molecular event, cellular process, and organism-scale phenotype. BEL is designed to represent discrete scientific findings and their relevant contextual information as qualitative causal relationships that can drive knowledge-based analytics. BEL enables biological interference by applications but furthermore is intended as an intuitive language of discourse for biologists. In such a way BEL is well aligned to the communications done in publications. The condensed representation of BEL statements and human as well as machine readability are great advantages of the BEL language.

3 Overview of basic concepts in BEL

BEL defines semantic triples that are stored in structured human readable BEL document files. A semantic triple is defined as a subject – predicate – object triple, where subject is always a BEL term, object either a BEL term or a BEL statement (recursive nature of BEL) and the predicate one of the BEL relationship types. A BEL term is composed of a BEL function, a corresponding entity and a referencing namespace. The two main classes of BEL terms define abundance of an entity (e.g., gene) or a biological process (e.g., disease).

Optionally, statements can be enriched by con-

[4] http://www.openbel.org/

[5] http://2013.bionlp-st.org/

[6] http://sbml.org

[7] https://sites.google.com/site/bionlpst2013/tasks/pathway-curation

[8] Comparison of BEL V1.0 and BioPAX Level3.pdf http://www.openbel.org/content/bel-lang-resource-documents

text information annotations like the evidence sentences, tissue type, species or cell line. Two annotation types are reserved, i.e., ‘Citation’ and ‘Evidence’. ‘Evidence’ should state the exact sentence that holds the statement’s information, where ‘Citation’ is the source of this knowledge.

Predefined namespaces cover a variety of biological entities: genes, proteins, chemicals, diseases and biological processes. For a complete definition of BEL we refer to the BEL Language documentation.

BEL Expression	**Explanation**
p(HGNC:AKT1)	Term: Protein Abundance function p(Ns:entity)
r(HGNC:AKT1)	Term: RNA Abundance function r(Ns:entity)
a(CHEBI:phosphoenolpyruvate)	Term: Chemical Abundance function a(Ns:entity)
p(HGNC:AKT1, sub(V,243,P))	Term: Protein Abundance function with substitution modification p(Ns:entity, sub(Aai,Pos,Aaj))
p(HGNC:AKT1, pmod(P,S,21))	Term: Protein Abundance function with phosphorylation modification p(Ns:entity,pmod(P,Aa,Pos))
kin (p(HGNC:AKT1))	Term: Protein Abundance function with kinase modification kin(p(Ns:entity))
complex (p(HGNC:CHUK), p(HGNC:IKBKB), p(HGNC:IKBKG))	Term: Complex Abundance function complex (p(Ns:entity)i,…, p(Ns:entity)n)
tloc(p(HGNC:EGFR), MESHCL: “Cell Membrane”, MESCL:Endosomes)	Term: Translocation function for Protein Abundance specifying the original and target location Tloc(p(Ns:entity), Ns:entity, Ns:entity)
deg(p(HGNC:AKT1))	Term: Degradation function for protein abundance deg(p(nNs:protein))
Reaction: rxn(reactants(a(CHEBI: phosphoenolpyruvate), a(CHEBI:ADP)), products (a(CHEBI:pyruvate), a(CHEBI:ATP)))	Statement: reaction expressing the transformation of products into reactants, each defined by a list of abundances rxn(reactants(a(Ns:entity)...), products(a(Ns:entity)...)
p(HGNC:IL6) -> r(HGNC:ENO1)	Statement: increase Term ->Term or Term -> Statement
p(HGNC:TNF) -| r(HGNC:NOS3)	Statement: decrease Term -|Term or Term -| Statement
p(HGNC:TNF) -- r(HGNC:NOS3)	Statement: association Term --Term or Term --Statement

Table 1: Example BEL terms and statements. Abbreviations: Ns=namespace, Aa=amino acid, Pos=position

In this work we focus mainly on protein-protein relationships (for simplification ‘protein’ refers to the corresponding gene, the RNA intermediate and the gene product itself[9]). Protein-protein relationships are a main focus of the BioNLP shared tasks and cover core relationships of BEL. An overview of possible statements is given in Table 1 and shortly described below. Protein entities are represented by BEL terms, consisting of the abundance function, the normalized entity and optionally modifications expressed as additional arguments within the abundance function:

BEL statement: *p(HGNC:AKT1, pmod(P, S, 21))*
Entity: *AKT1*
Namespace: *HGNC*
Optional modification: *pmod(P,S,21)*

The used namespace denotes the approved symbol of HUGO Gene Nomenclature Committee[10]. An overview of currently used namespaces is given at the OpenBEL portal. The pmod() function explicitly denotes the modification type (here *P*=phosphorylation), the 1-letter code for the corresponding amino acid (*S*=Serin) and the position in the protein sequence. Other modifications are represented with different codes, e.g., M=methylation or U=ubiquitination.

BEL terms may contain protein activity information such as kinase or transcription factor activity or certain functions like complex, degradation, translocation or reaction in addition.

```
SET Citation = {"PubMed","Cell","16962653","2006-
10-07","Jacinto E|Facchinetti V|Liu D|Soto N|Wei
S|Jung SY|Huang Q|Qin J|Su B",""}
SET Cell = "Fibroblasts”
SET Species = "10090"
SET Evidence = "We next examined the Akt T-loop
Thr308 phosphorylation in wild-type and SIN1−/−
cells. We found that although Ser473 phosphorylation
was completely abolished in the SIN1−/− cells, Thr308
phosphorylation of Akt was not blocked (Figure 3A)."

p(MGI:Mapkap1) -> p(MGI:Akt1,pmod(P,S,473))
p(MGI:Mapkap1) causesNoChange
p(MGI:Akt1,pmod(P,T,308))
```

Figure 1: Example of enriched BEL Statement

By default (but not mandatory) ‘Evidence’ and ‘Citation’ annotations are provided for each

[9] according to BioNLP shared tasks annotations

[10] http://www.genenames.org/data/hgnc_data.php?hgnc_id=391

statement. In case of extraction from literature the reference source and the evidence sentences are given. Alternative evidences may be derived from tables, figures, supplementary material or other knowledge sources. Optionally, the BEL statements can be annotated with specified information about experimental methods, the biological system in which the facts are represented, or even information in which part of the full text the evidence has been found. An example of such a BEL statement from a small sample set at the OpenBEL portal[11] is shown in Figure 1. Such detailed information from literature, in combination with the BEL statements, could serve as ideal source for the generation of training data for text mining purposes to facilitate the development of future automated extraction algorithms.

4 Analysis of basic concepts in the BioNLP shared task annotations

In the main BioNLP shared task (GE12) nine event types are defined (cf Table 2). 'Gene expression', 'Transcription', 'Protein catabolism', 'Phosphorylation' and 'Localization' are simple events, having one protein as *Theme* argument.

Event	Primary Arg.	Secondary Arg.
Gene Expression	Theme(Protein)	
Transcription	Theme(Protein)	
Protein Catabolism	Theme(Protein)	
Phosphorylation	Theme(Protein)	Site
Localization	Theme(Protein)	AtLoc, ToLoc
Binding	Theme(Protein)+	Site+
Regulation, Positive Regulation, Negative Regulation	Theme(Protein/Event) ,Cause(Protein/Event)	Cause, Site, CSite

Table 2: Event types defined in the BioNLP competitions (adapted from (Kim et al., 2012). A '+' sign indicates multiple occurrences allowed.

Events 'Phosphorylation' and 'Localization' may have additional secondary arguments, like the phosphorylation site or the localization arguments ToLoc and AtLoc. 'Binding' events can have an arbitrary number of proteins as *Themes*. Events 'Positive regulation', 'Negative regulation' and 'Regulation' are *Regulation Events* and have a primary *Theme* argument and an optional *Cause* argument, both being either a protein or an event. The trigger is always the textual representation of the entities. Table 3 depicts an example annotation for the following sentences[13]:

S1) E1-4: "RFLAT-1: a new zinc finger transcription factor that activates RANTES gene expression in T lymphocytes."

S2) E5-9: "In this study we hypothesized that the phosphorylation of TRAF2 inhibits binding to the CD40 cytoplasmic domain."

ID	Theme Type	Trigger	Theme	Cause
T1	Protein	RFLAT-1		
T2	Protein	RANTES		
E3	Gene Expression	gene expression	T2	
E4	Positive Regulation	activates	E3	T1
T5	Protein	TRAF-2		
T6	Protein	CD40		
E7	Phosphorylation	phosphorylation	T5	
E8	Binding	binding	T6	T5
E9	Negative Regulation	inhibits	E8	E7

Table 3: Example BioNLP 09 shared task annotation. The gene/protein entities with the Ids T1, T2, T5, and T6 were already provided. The task was to detect the events E3, E4, E7, E8 and E9.

5 Syntactic mapping from BioNLP annotation to BEL statements

For mapping of the BEL statements and the output of the BioNLP shared tasks systems we compared the training data for the GENIA BioNLP task with the BEL statements found in the small corpus at the OpenBEL website. The BioNLP shared task provides no normalization of the entities to namespaces. Since we are mainly interested in the transformation of the event, we ignore the normalization aspect in the conversion process. For most Shared Task events we could

[11] https://github.com/OpenBEL/openbel-framework-resources/blob/master/knowledge/small_corpus.bel

[12] https://sites.google.com/site/bionlpst/home/genia-event-extraction-genia

[13] Examples taken from http://www.nactem.ac.uk/tsujii/GENIA/SharedTask/detail.shtml

generate BEL Terms which are summarized with the rule set in Table 4 and Table 5.
Standard translation for all protein Themes is protein abundance *p(namespace:entity)*. In a later network generation step within the BEL framework RNA abundance and gene abundance are added automatically to the network of statements for all protein abundances. Due to this reason, we only consider RNA or gene abundance if we detect strong evidences for those states. For Gene_expression, the protein abundance is only converted to RNA abundance (r(namespace:entity)) if the trigger word is 'gene expression'.

1.1	GeneExpression(Theme(protein)) → p(Ns:protein) If the GeneExpression trigger word is stemmed to 'express'
1.2	GeneExpression(Theme(protein)) → r(Ns:protein) For all other GeneExpression trigger words.
2	Transcription(Theme(protein)) → r(Ns:entity)
3	Phosphorylation(Theme(protein), <Site>) → p(Ns:protein, <pmod(P,*Aa, Pos*)>)
4	ProteinCatabolism(Theme(protein))→ deg(p(Ns:protein))
5.1	Localization(Theme(protein)) → *sec* (p(Ns:protein)) If the Localization trigger is stemmed to 'secrete'
5.2	Localization(Theme(protein),AtLoc) → *surf*(p(Ns:protein)) If the Localization trigger is stemmed to 'express' and If AtLoc is 'cell surface' or 'surface'
5.3	Localization(Theme(protein),AtLoc, ToLoc) → *tloc* (p(Ns:protein),Ns:AtLoc,Ns:ToLoc) In BEL statements it is necessary to have AtLoc and ToLoc; for some cases the missing information can be inferred otherwise artificial location information is given.
6	Binding(Theme(protein)+,Site+) → complex(p(ns:protein),+) The site information will be ignored.

Table 4: Rule set 1 to map BioNLP annotations to BEL statements.

If the trigger word 'expression' is used, both RNA and protein expression might be meant by the authors, hence we keep the protein abundance in those cases. Similarly for Transcription, the abundance is changed to RNA abundance. All complexes are translated to protein abundance and chemical names are directly translated into abundance (a(ns:chemical names)). Protein modification events such as Phosphorylation can be directly converted to BEL terms. The different modification events are translated to a single letter code in BEL. If the position information is given in the site expression it can directly be converted to the amino acid single letter code (Aa) and the position information (Pos). For the simple events Protein degradation and Binding, the translation is straightforward given their similar representation. The site information of the Binding event is omitted in the BEL statement conversion. It would only be included if there is an experiment showing that a mutation of the site would lead to a suppression of the complex building.

In the case of 'Localization', depending on the localisation trigger different BEL functions are possible. Given the localization trigger 'secrete' the BEL annotation is converted to the secretion (sec) function. If trigger words 'surface' or 'cell surface' are identified, the cellSurface (surf) function is assigned. For other Atloc and ToLoc triggers the function translocation (tloc) is used. This function always needs two arguments of location. If one of the arguments (AtLoc or ToLoc) is missing, a general annotation of MESHCL:"Intracellular Space" is proposed as unknown intracellular location.

Activity status like gtp(p(protein)), kin (p(protein)), tscript(p(protein)), cat(p(protein)), phos(p(protein)) are often found in the BEL example corpus. This information might be partly inferred through the evidence information. In the first example sentence from Table 2, RFLAT-1 might be directly translated into tscript (p(RFLAT-1)). In other cases if a protein phosphorylates another protein directly, the kin(p(protein)) annotation can be added as well. However, in most cases the information cannot directly be inferred from the sentences (cf. Figure 1). The annotators obviously use their background knowledge to include this information. In the actual status of the Shared Task to BEL conversion we omitted those functions.

Looking at the rule-set for transferring Shared-Task events to BEL statements, it is observed that for most events (six out of nine) only *BEL terms* are generated, i.e., only the left or right hand side of a complete statement. Three rules generate complete BEL statements out of the following events: *Regulation*, *Positive Regulation* and *Negative Regulation*. Analysis of the distribution of Events in Shared-Tasked training set (BioNLP ST 2011) reveals that approximately half of the events are Regulation events and thus, could lead to a set of complete statements. In Table 5, we describe the rules which generate complete BEL statements.

7	PositiveRegulation(Theme(Protein/Event), Cause(Protein/Event)) → p(ns:protein)/B(Event) -> p(ns:protein)/B(Event)
8	NegativeRegulation(Theme(Protein/Event), Cause(Protein/Event)) → p(ns:protein)/B(Event) -| p(ns:protein)/B(Event)
9	Regulation(Theme(Protein/Event), Cause(Protein/Event)) → p(ns:protein)/B(Event) -- p(ns:protein)/B(Event)

Table 5: Rule set 2 to map BioNLP annotations to BEL statements.

For all 'Regulation' events the *Theme* is translated to the object of the BEL statement and might be a protein or another BEL statement (B(Event)). The *Cause* is integrated as subject within the statement and can be a protein or a statement. All 'Positive Regulation' events in the Shared Task annotations are converted to 'increase' statements of BEL. We do not differentiate between 'increase' and 'directly increase' in the conversion process. Similarly, all 'Negative Regulation' events are converted to a 'decrease' statement ignoring 'directly decrease'. In the BEL annotations those two statement groups are the most frequent statements in both corpora. In the Shared Tasks relations we have the additional relation Regulation. There is no directly corresponding BEL relation for a general regulation event, since it restricts the impact for causal reasoning. The event which has the most similar meaning is the statement 'association'. It is used for associations of proteins but also for associations of proteins and diseases when no further information is available in the text. The additional annotations Site and CSite are currently ignored since there is no structure in BEL to include this information directly.

In all three regulation events the Cause is an optional argument and might be missing. Out of the 7574 regulation events 2152 events contain a cause and thus can be converted to a complete BEL statements. For all other events the left hand side of the statement is missing.

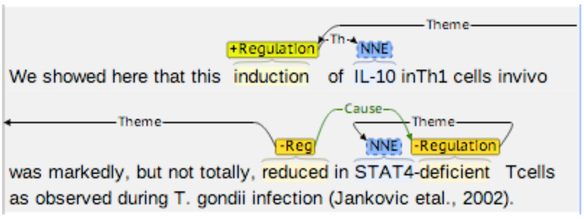
For obtaining an overview of the conversion process we converted the event annotations from the GENIA training corpus to BEL statements (all relations containing a speculation or a negation were omitted). The automatically generated BEL documents were checked for syntactical errors with the OpenBEL framework parser and validator. Several adaptations were necessary in the automatic conversion process to generate syntactically correct BEL statements.

Since we have no namespaces available we designed an artificial namespace to generate correct statements. Furthermore incomplete statements with missing subjects (Causes) were not accepted by the BEL framework. An example of such an incomplete BEL statement is the following (converted form the shared task annotation depicted in Figure 2):

-| p(BioNLP:STAT4) -| p(BioNLP:IL10)

For all missing *Causes* we included an artificial *Cause* resulting in the following statement for the given example:

p(BioNLP:FIXME)-| p(BioNLP:STAT4) -| p(BioNLP:IL10)



We showed here that this induction of IL-10 inTh1 cells invivo was markedly, but not totally, reduced in STAT4-deficient Tcells as observed during T. gondii infection (Jankovic etal., 2002).

Figure 2: An example sentence from BioNLP-ST 2011 GE train corpus, visualized using brat. [14]

Overall 5333 BEL statements were generated resulting in 588 full statements, 3057 incomplete statements (where the CAUSE is missing and FIXME was introduced) and 1688 BEL terms without any relation. Remaining syntactic errors were caused through BEL statements containing more than two relations (118 statements), which could not be handled by the BEL framework. A first version of the converted corpus is available under: http://www.scai.fraunhofer.de/ge2011-to-bel.html.

6 Preliminary comparison of converted statements with BEL knowledge resources

In the BioNLP shared tasks all possible events that fulfill the guidelines are annotated. In real life use-cases irrelevant or unproven interactions are omitted and biological experts extract BEL statements when they are in focus of their interest. Furthermore experimental evidence for the relation should be should be given in the text.

[14] http://brat.nlplab.org

In addition biologists are able to do a semantic interpretation of the experimental results and generate inferred statements. To find solutions for semantic interpretation for a number of incomplete statements in the direct conversion for the BioNLP-ST annotations we compared sentences such as annotated in figure 2 with evidence sentences in the BEL sample set. In the following examples we show how an expert curator conversely infers BEL statements by interpreting experiment readouts.

Example 1:
Evidence = "PI 3- kinase/PKCξ, but not PI 3-kinase/Akt signaling pathway, is inhibited in IRS-2-deficient brown adipocytes upon insulin stimulation"

p(HGNC:IRS2)-> kinase(p(HGNC:PRKCZ))
p(HGNC:IRS2) causesNoChange kin(p(HGNC:AKT1))

Example 2:
Evidence = "transient transfection of primary brown adipocytes with a dominant negative form of p21 Ras completely abolished insulin-induced UCP-1-CAT transactivation."

p(PFH:"RAS Family") -> (p(HGNC:INS) -> r(HGNC:UCP1))

Example 3:
Evidence = "We next examined the Akt T-loop Thr308 phosphorylation in wild-type and SIN1−/− cells. We found that Thr308 phosphorylation was completely abolished in the SIN1−/− cells."

p(MGI:Mapkap1) -> p(MGI:Akt1,pmod(P,T,308))

The examples given above demonstrate a standard experimental setting. In most cases the functionality of a gene is abolished and the effect (e.g. increase, decrease or no effect) on the corresponding interaction targets is observed. Sometimes, observed effects are compared to cell systems where the normal form (wild type or control) is transfected as well (cf. Example 3).

All examples share the readout: The BEL statement is not describing the experiment (given in the sentence), but the observed implication inferred from the experiment (cf. Example 2). Instead of encoding that a dysfunctional p21 RAS leads to an abolishment of insulin induced UCP1 transactivation, the final BEL statement represents the resulting implication, i.e. wild-type p21 RAS increases INS, which subsequently increases UCP1:

p(PFH:"RAS Family") -> (p(HGNC:INS) -> r(HGNC:UCP1))

Similarly, in Example 3 from the abolishment of a function, the converse argument is derived, i.e. Mapkap 1 increases the phosphorylation of Akt1 at T308. This example shows another main issue in deriving BEL statements: two or more sentences are needed to get all information necessary to create a valid BEL statement. Human curators use multiple sentences as evidence and do additional interpretation of the provided information. In Example 3, the AKT phosphorylation is given in the first sentence and the phosphorylation event is given in the following sentence only in referring to the site and not to the protein. BioNLP-ST already includes annotation spanning several sentences but interpretation and merging of those annotations is not trivial. To complete such statements two different relations have to be combined and that is true for many modification relations. Especially in the case of phosphorylation, which is a regular activating signal in kinase pathways, we need solutions including information from different sentences. The BEL corpus has a high number of phosphorylation events and can serve as a base for the generation of further training data.

Another commonly observed experiment uses luciferase and CAT vectors. Those systems are used to analyze transcriptional activity of promoters in dependence of stimuli. The result of such an experiment is oftentimes given only as a relation to CAT or luciferase like in the following example:

Example 4:
Evidence = "introduction of miR-145, but not miR-143, with the luciferase vector in Cos cells resulted in relief of the repression and an ~150-fold increase in luciferase activity compared to the CMV-luciferase- Myocd 3' UTR-luciferase vector alone."

miR(HGNC:MIR145) -> p(HGNC:MYOCD)
miR(HGNC:MIR143) causesNoChange p(HGNC:MYOCD)

BioNLP shared task annotation would capture positive regulation of luciferase activity with the cause miR-145. The derived statement however does not state an abundance function for luciferase but the originally tested protein (indirectly via its promotor) i.e., Myocd. Here, the inserted promoter information is given at the end of the sentence, although it is often provided in a separate sentence.

The second BEL statement in Example 4 provides another relation type, which is not directly captured by the shared task annotations. Negative results are annotated in BEL statements with the relation causesNoChange and are valuable relations in causal reasoning. They might be interpreted using the negation annotation in shared task to capture this type of event.

Those examples are only a few out of numerous others. For the development of suitable systems, annotated training corpora are crucial. The BEL documents might be a good starting point to generate further training corpora containing a high number of such evidence examples. However, the conversion of the BEL statements to BioNLP shared task annotation is not trivial, since position information is completely missing. Nevertheless, it might reduce the annotation effort, give good examples and serve as a basis for biological interpretation of the relations. For initial automatic systems it might be even sufficient to offer such experimental evidence sentences in addition to the extracted relations to users.

7 Discussion and Conclusions

Generally, a syntactic conversion of BioNLP shared task annotations to BEL terms and statements is possible and in most cases without information loss. Tools developed or adapted for the BioNLP shared task are principally suited for the generation of causal BEL networks. However, the analysis of the automatically converted BEL statements from the BioNLP shared tasks shows that in a number of cases incomplete BEL statements were generated. Part of the reason is the need for an additional interpretation layer that would help in generating biologically meaningful statements. Another reason for the failure to extract full statements is the distribution of the relation over more than one sentence.

The properties of BEL statements and the additional information coded in the BEL documents represent a valuable resource for generating further training data for the development of more real-world oriented systems. Unfortunately, the information of the BEL documents cannot directly be converted back to textual annotation. The main reason is that the position information of entities within the relation is missing. Reverse engineering is also challenging because the trigger words are not given. Furthermore, normalization to namespaces used in BEL statements makes the direct mapping difficult.

Nevertheless, the text mining community can learn from the BEL documents what are relevant statements for causal reasoning and from which evidence sentences humans extract the information. The example BEL statements given show that humans use a number of experimental systems such as inactive versions of proteins or reporter genes to prove existing relationships. It might be a realistic task to use BEL documents as a starting point to generate training corpora for the automatic classification of such sentences and for information extraction systems to extract relations from those sentences. For some relations like the phosphorylation or the reporter genes, we might be even able to extract relations over sentences when enough training data is available.

Another problem not tackled by the BioNLP shared tasks is the mapping to the name spaces. There are already systems available combining BioNLP based relation extraction systems and named entity recognition (NER) systems allowing for normalization and (eg. Björne et al., 2012 and Van Landeghem et al., 2013). Future systems have to combine relation extraction and NER systems allowing for normalization. Gene and protein names have already been in the focus of the BioCreative assessments during the last years (cf. Morgan et al., 2008 and Lu et al., 2011). In addition, chemical entities are coming more and more into the focus of the community (e.g., in the BioCreative 2013 task[15]). In the examples from the BEL corpus we see additional problems coming from the area of engineered genes. Name variants are often used (e.g., Sin-/- or CMV-luciferase- Myocd 3' UTR-luciferase), which causes further problems in the normalization task.

Bridging the BEL and the BioNLP-ST community offers benefits for both sides. The BioNLP shared tasks are a considerable start for the automatic generation of causal networks. Moreover, already available BEL documents can support the generation of the huge amount of additional training data, which is necessary for further relation extraction development.

[15] http://www.biocreative.org/events/biocreative-iv/CFP/

Acknowledgments

We would like to thank Natalie Catlett and Ted Slater of Selventa for providing their time and expertise in helping us understand BEL. We acknowledge support of our research from Philip Morris International.

References

Jari Björne, Sofie Van Landeghem, Sampo Pyysalo, Tomoko Ohta, Filip Ginter, Yves Van de Peer, Sophia Ananiadou and Tapio Salakoski. 2012. PubMed-Scale Event Extraction for Post-Translational Modifications, Epigenetics and Protein Structural Relations. *Proceedings of BioNLP 2012, 82-90*

Leonid Chindelevitch, Daniel Ziemek, Ahmed Enayetallah, Ranjit Randhawa, Ben Sidders, Christoph Brockel and Enoch Huang. 2012. Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics. 28(8):1114-21.*

Emek Demir et al. 2010. The BioPAX community standard for pathway data sharing. *Nature biotechnology , 28(9):935–942.*

Chia-Ling Huang, John Lamb, Leonid Chindelevitch, Jarek Kostrowicki, Justin Guinney, Charles DeLisi and Daniel Ziemek. 2012. Correlation set analysis: detecting active regulators in disease populations using prior causal knowledge. *BMC Bioinformatics. 2012 13:46.*

Michael Hucka et al. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics , 19(4):524–531.*

Jim-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi and Akinori Yonezawa. 2012. The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011. *BMC Bioinformatics. 13 Suppl 11:S1.*

Zhiyong Lu et al. 2011. The gene normalization task in BioCreative III. *BMC Bioinformatics , 12(Suppl 8):S2.*

Alexander A Morgan et al. 2008. Overview of BioCreative II gene normalization. *Genome Biol. 9 Suppl 2:S3.*

Tomoko Ohta,_ Sampo Pyysalo, Sophia Ananiadou and Jun'ichi Tsujii. 2011. Pathway Curation Support as an Information Extraction Task. *Proceedings of the Fourth International Symposium on Languages in Biology and Medicine (LBM 2011).*

Tomoko Ohta, Sampo Pyysalo and_Jun'ichi Tsujii. 2011. From Pathways to Biomolecular Events: Opportunities and Challenges. *Proceedings of the 2011 Workshop on Biomedical Natural Language Processing , ACL-HLT 2011, pages 105–113.*

Walter K. Schlage, et al. 2011. A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. *BMC Syst Biol. 5:168.*

Ted Slater and Diana H. Song. 2012. Saved by the BEL: ringing in a common language for the life sciences. *Drug Discovery World Fall 2012 75:80*

Selventa 2012 Reverse Causal Reasoning Methods Whitepaper http://www.selventa.com/publications/white-papers

Sofia Van Landeghem, Jari Björne, Chih H Wei, Kai Hakala , Sampo Pyysalo, Sophia Ananiadou,

Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. Large-Scale Event Extraction from Literature with Multi-Level Gene Normalization. *PLoS ONE 8(4): e55814.*