Track 4 Overview: Extraction of Causal Network Information in Biological Expression Language (BEL)

Fluck, Juliane ; Madan, Sumit ; Ellendorff, Tilia Renate ; Mevissen, Theo ; Clematide, Simon ; van der Lek, Adrian ; Rinaldi, Fabio

Track 4 Overview: Extraction of Causal Network Information in Biological Expression Language (BEL)

Juliane Fluck[1], Sumit Madan[1], Tilia Renate Ellendorff[2], Theo Mevissen[1], Simon Clematide[2], Adrian van der Lek[2] and Fabio Rinaldi[2]

[1]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany.
[2] Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland.

juliane.fluck@scai.fraunhofer.de; sumit.madan@scai.fraunhofer.de; ellendorff@cl.uzh.ch; theo.mevissen@scai.fraunhofer.de; siclemat@cl.uzh.ch; adrian.vanderlek@uzh.ch; fabio.rinaldi@uzh.ch

Abstract. Automatic extraction of biological network information is one of the most desired and most complex tasks in biological text mining. The BioCreative track 4 provides training data and an evaluation environment for the extraction of causal relationships in Biological Expression Language (BEL). BEL is a modeling language that is easily editable by humans or by automatic systems and can express causal relationships of different levels of granularity. Protein-protein relations can be expressed in BEL as well as relations between biological processes and disease stages. To extract BEL information automatically, named entity recognition and normalization to defined name spaces are necessary. Furthermore, relations extracted from text have to be transformed into correct BEL syntax. The track provided training and evaluation for two complementary task: Given a sentence extract all BEL statements and given a BEL statement propose up to 10 evidence sentences from the literature.

Keywords: Networks; Causal Relationships; Information Extraction; NLP

1 Task Overview

Biological networks with a structured syntax are a powerful way of representing biological information and knowledge. Well-known examples of methods to formally represent biological networks are the Systems Biology Markup Language (SBML) [1], the Biological pathway exchange language (BioPAX) [2] and the Biological Expression Language[1] (BEL) [3]. These approaches are not only designed for the representation of biological events, but they are also intended to support downstream computational applications. In particular, BEL is gaining ground as the de-facto standard for systems biology applications because it combines the power of a formalized repre-

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

sentation language with a relatively simple syntax that allows an easy interpretation of BEL statement by a trained domain expert. BEL has originally been developed by Selventa, a personalised healthcare organisation, and used primarily in industrial research for large data interpretation. However, in 2013, BEL became a Linux foundation collaborative project.

As part of an on-going product assessment program, the sbvIMPROVER initiative is supporting the manual curation and expansion of biological networks related to human lung disease [4–7]. They organized a large-scale crowdsourcing verification approach for the verification of these biological networks, called Network Verification Challenge (NVC) [8]. This initiative aims to provide a measure of quality control for systems based research, supporting the verification of methods and concepts in this domain. The NVC supports community-based verification and extension of biological relationships based on peer-reviewed literature evidence. At present, 50 biological networks have been curated, all available in BEL format, with supporting evidence in form of a sentence or section and a PubMed identifier.

Based on the data provided and validated through the sbvIMPROVER NVC, we prepared training and test corpora to initiate novel approaches aiming at relation extraction and automated construction of network elements. The goal is to assess the utility of such tools either for the automated annotation and network expansion, or their suitability as supporting tools for assisted curation. The challenge is organized into two tasks evaluating the complementary aspects of the problem:

Task 1: Given textual evidence for a BEL statement, generate the corresponding BEL statement.

Task 2: Given a BEL statement, provide at most 10 additional evidence sentences.

In the following a short overview of the Biological Expression Language BEL is given and the preparation of training data is described. Furthermore, the evaluation criteria for the participating systems are explained and their results are shown.

2 Biological Expression Language

The Biological Expression Language (BEL) is designed to represent scientific findings in the field of the life sciences in a form that is not only computable but also easily editable by humans. The findings are captured through causal and correlative relationships between entities in the format of BEL statements. Publication references are provided as supporting information for each statement. Most BEL statements represent relationships between one BEL Term and another BEL term or BEL statement. Example BEL statements are shown in Figure 1. The statements typically encode a semantic triple (subject, relationship type and object). These triples represent an assertion of a relationship between the subject and object. For track 4, a focus was made on the causal relationships shown in Table 1.

Training set entry provided to participants:

Training.sentence entry:

SEN:10000032	PMID:10075927	Fas stimulation of Jurkat cells is known to induce p38 kinase and we find a pronounced increase in Rb phosphorylation within 30 min of Fas stimulation.

Training.BEL entry:

SEN:10000032	cat(p(HGNC:FAS)) increases p(HGNC:RB1,pmod(P))	BEL:20006082
SEN:10000032	p(HGNC:RB1,pmod(P)) directlyDecreases tscript(p(HGNC:RB1))	BEL:20011414
SEN:10000032	cat(p(HGNC:FAS)) increases kin(p(HGNC:MAPK14))	BEL:20029764
SEN:10000032	cat(p(HGNC:FAS)) decreases tscript(p(HGNC:RB1))	BEL:20029794

Fig. 1. Training data example

Relationship – long form	Short form	Example
decreases	-|	a(CHEBI:"brefeldin A") -| p(HGNC:SCOC)
directlyDecreases[1]	=|	p(HGNC:TIMP1) =| act(p(HGNC:MMP9))
increases	->	p(MGI:Bmp4) -> p(MGI:Acta2)
directlyIncreases[2]	=>	p(HGNC:VEGFA) => act(p(HGNC:KDR))

[1] decreases is accepted instead of directlyDecreases;
[2] increases accepted instead of directlyIncreases

Table 1. Relationships part of Track 4

The specifications of BEL allow for an easy adaptation of external vocabularies and ontologies. BEL adopts a concept of namespaces to disambiguate references to entities. By applying namespace prefixes a user can establish references to elements of the specific vocabulary. Currently, BEL offers more than 20 different namespaces. For simplification purposes the dataset used in track 4 was restricted to a selection of 6 namespaces (c.f. Table 2). Different namespaces have different abundance and process functions associated. BEL terms are formed using these BEL functions together with the namespaces and the associated identifiers. Each BEL term represents either a biological process or the abundance of an entity. An overview of short and long function names associated to namespaces can be found in Table 2. In order to find equivalences between the entities of different namespaces, a range of equivalence resources are provided at the OpenBEL website[2]. During the compilation process of the BEL framework these equivalences are incorporated. Therefore, all gene/protein

[2] https://github.com/OpenBEL/openbel-framework-resources/tree/latest/equivalence

namespaces were treated as equivalent in the evaluation. Furthermore, orthologous entities were accepted as true positive hits.

Name space	Function Long form	Function Short form	BEL Term Example
HGNC	geneAbundance(), rnaAbundance(), microRNAAbundance(), **proteinAbundance()**	g(), r(), m(), **p()**[1]	p(HGNC:MAPK14)
MGI	Similar to HGNC	Similar to HGNC	p(MGI:Mapk14)
EGID	Similar to HGNC	Similar to HGNC	p(EGID:1432)
GOBP	**biologicalProcess()**	**bp()**	bp(GOBP:"cell proliferation")
MESHD	**pathology()**	**path()**	path(MESHD:Hyperoxia)
CHEBI	**abundance()**	**a()**	a(CHEBI: lipopolysaccharide)

[1] p() was accepted instead of g(), r(), m()

Table 2. Overview of Track 4 namespaces and associated functions

Information about the state (e.g. transformation, translocation or molecular activity) in which entities are found, is encoded as functions, which take BEL terms as arguments. An overview of selected functions for the task is provided in Table 3.

3 Preparation of Training Data

BEL networks provided by the Improver Network Verification challenge were used as a starting point for the generation of training and test corpora. A part of these BEL networks is publically available[3] [9]. Those statements were mainly extracted from abstracts or full text papers. The following selection criteria were defined for the training corpus:

- Statement is not inferred automatically from the compiler
- Statement is associated with a PubMed Citation
- Statement evidence (summary text) is associated with fewer than 5 statements in total to avoid statements from tables
- Statement evidence has a length between 36 – 425 characters to focus on evidences based on one or two sentences

[3] Can be downloaded at http://www.causalbionet.com/

Function	Function Type	Example
complex() *complexAbundance()*	Abundances	(complex(p(MGI:Itga8),p(MGI:Itgb1))) -> bp(GOBP:”cell adhesion”)
pmod() *proteinModification()*	Modifications	p(MGI:Cav1,pmod(P)) -> a(CHEBI:”nitric oxide”)
deg() *degradation()*	Transformations	p(MGI:Lyve1) -> deg(a(CHEBI:”hyaluronic acid”))
tloc() *translocation()*	Transformations	a(CHEBI:”brefeldin A”) -> tloc(p(MGI:Stk16))
act() *molecularActivity()*	Activities	complex(p(MGI:Cckbr),p(MGI:Gast)) -> act(p(MGI:Prkd1))

Table 3. Overview of selected functions

In order to reduce the complexity of the BioCreative task, we selected statements containing only a specific subset of entity classes, relationship types and functions (cf. chapter 2 for a description of the selected categories). Furthermore, context annotations were ignored completely within this task. As a result, the following filter criteria were automatically applied in a second step:

- Statement relationship is increases, decreases, directlyIncreases or directlyDecreases
- Statement contains only HGNC, MGI, EGID, MESHD, CHEBI or GOBP name space entities as subject or object terms
- Statements with less or equal than 4 entities
- Statements without the functions *composite()or rxn()*

The resulting corpus for training and test set generation contained 12,268 statements. From this corpus, a set of 6,353 sentences accompanied with 11,066 statements were published as training data. The file *training.sentence* contains the sentence ID, the PubMed Identifier (PMID) and the evidence sentence. The file *training.BEL* contains the sentence ID and the BEL statement and a BEL-ID. Examples of training set entries are shown in Figure 1.

As can be seen from the given example, not all statements can be extracted from the evidence sentence. Statement BEL:20011414 and BEL:20029794 can only be inferred from background knowledge or from other sentences of the same publication. This is true for many statements and even more for the activity functions such as *cat()*. For these reasons, a sample set was published. An annotator checked the corpus sentences to approve that they contain the entity mentioned as well as their relationship. Analysis of evaluation results based on the sample set showed that relations can be coded in different ways. Often, experts chose only one way to

annotate the relation. To make a better evaluation feasible the test corpus BEL statement set was extended. All possible relations that could be derived from the sentence and were based on the defined name spaces were added as BEL statements. This led to 202 statements in 105 sentences. An example of such an extension is given in Figure 2. This was a deviation from the sample set and training data and might lead to a performance bias between training and test data.

Regarding the task 2 test set, we verified that at least one PubMed sentence could be assigned to the provided BEL statements. In addition, those statements could not be generated from the task 1 sentence set. Furthermore, an annotator approved the correctness of the statements. Overall, the task 1 test set contained 105 sentences and the task 2 test set 100 BEL statements.

Table 4 gives an overview of the different items in the training and test set. There is a dominant category type on each level in the training set: 87% of the terms are proteins, 69% of the functions are activations, and 73% of the relations express an increase. Similar proportions apply to the test set, except for the function level where activation covers only 46% of all cases.

4 Supporting Resources

The participants were provided with a range of supporting resources and a comprehensive documentation[4], containing a description of the format and detailed explanation of the evaluation process. The evaluation on the different levels of a single BEL statement was illustrated using a set of concrete example submissions as reference. Additionally, an evaluation interface[5] was provided for the participants to test their generated statements during the development phase.

Further supporting resources included the BEL statements from the training and sample set in BioC format, which we generated automatically using a converter based on the official ruby-based BEL parser[6] and an open-source BioC ruby module[7] [10]. A tab-separated format that contains all fragments of the BEL statements (terms, functions and relations) was automatically generated from the sample and training set, using the same BEL parser mentioned above. These were provided to the participants as supporting material.

Finally, graph visualizations were generated based on the BioC format of the statements. An example for such visualization can be seen in Figure 3.

[4] http://wiki.openbel.org/display/BIOC/Biocreative+Home
[5] http://bio-eval.scai.fraunhofer.de/cgi-bin/General_server.rc
[6] http://www.openbel.org/tags/bel-parser-belrb
[7] https://github.com/dongseop/simple_bioc

Test set entry provided to participants:

SEN:10004710 PMID:15671176 More importantly, the Dnmt1 knockdown blocked the methionine-induced reelin and GAD67 mRNA down-regulation.

BEL statements in original corpus:

a(CHEBI:methionine) decreases r(MGI:Reln)
a(CHEBI:methionine) decreases r(MGI:Gad1)
p(MGI:Dnmt1) increases (a(CHEBI:methionine) decreases r(MGI:Gad1))
p(MGI:Dnmt1) increases (a(CHEBI:methionine) decreases r(MGI:Reln))

Added to GOLD standard:

p(MGI:Dnmt1) decreases r(MGI:Gad1)
p(MGI:Dnmt1) decreases r(MGI:Reln)

Fig. 2. Extension of test data

Term			**Function**			**Relation**		
Type	*Train*	*Test*	*Type*	*Train*	*Test*	*Type*	*Train*	*Test*
p	19918	346	Act	6332	36	increases	8112	155
a	1927	37	Pmod	1411	9	decreases	2956	53
bp	877	31	complex	750	15			
path	244	15	Tloc	406	13			
			deg	205	6			
			sub	23				
			trunc	6				

Table 4. Distribution of term, function and relationship types in the training and test set

5 Evaluation Criteria

Extraction of relationships and their coding in BEL is a complex task due to the different entity types, relationship types and functions. Furthermore, not all information coded in the expert-generated BEL statements can be found in the sentences provided as training data. Therefore, we simplified the relationships further and provided a cascade model for evaluation.

A detailed overview of all simplifications was provided online[8]. Firstly, HGNC or EntrezGene identifiers are accepted for the same statement; orthologous identifiers are accepted as well (for HGNC, MGI or EGID namespaces). Secondly, the evaluation does not differentiate between *increases* and *directlyIncreases* and *decreases* and *directlyDecreases* relationship types. Thirdly, various activity functions such as *kin(), tscript()* and *cat()* are transformed into *act()* and only this function is evaluated. Final-

[8] http://wiki.openbel.org/display/BIOC/All+Functions+Evaluation+Overview

ly, the modification function *pmod()* and the translation function *tloc()* were reduced in their number of arguments. *pmod(P)* is evaluated without the position and amino acid information and the *tloc()* function is evaluated without information of the location.

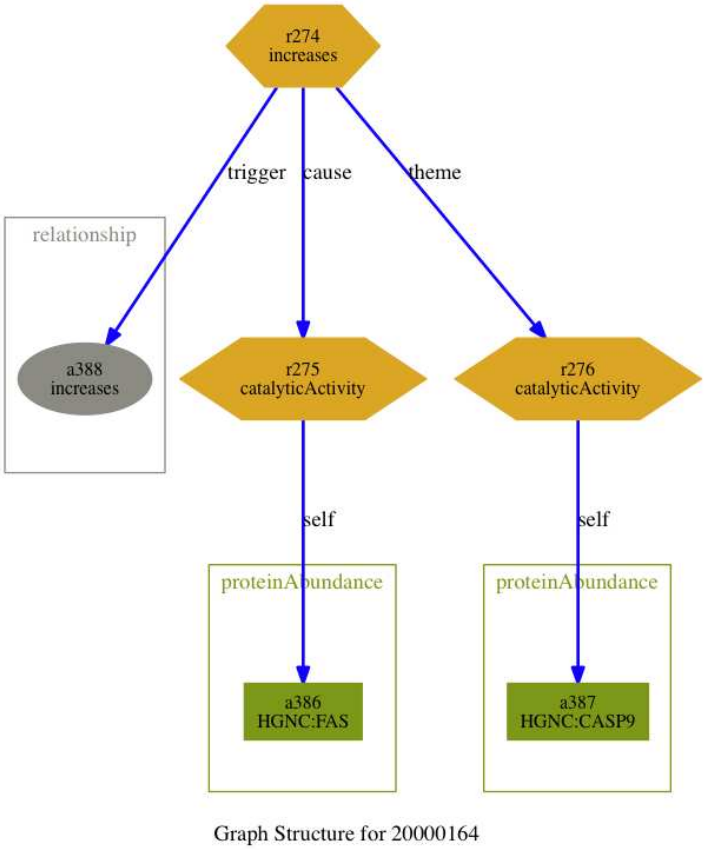

Fig. 3. Visualization of the BEL statement "cat(p(HGNC:FAS)) increases cat(p(HGNC:CASP9))" derived from the sentence "we demonstrate that two of the early events after Fas ligation are the release of cytochrome c from the mitochondria and activation of caspase-9"

In the cascade evaluation model, different levels of performance are evaluated. Since we use a formal language, BEL statements or fragments must be syntactically correct to be accepted for evaluation. Therefore, BEL terms (representing entities) must be complete and in a correct format, otherwise a submission will not be evaluated. Using false abundance or a process function, false name spaces or false variants of terms (e.g. false case or missing quotes for multi word terms) leads to false evaluation

results. The evaluation web service was provided to the participants to check for formal correctness during the development phase. This service could be used during the training phase of track 4 and submitted statements were evaluated based on the sample set.

Placeholder entities and relationship types were introduced to allow the submission of incomplete information. Instead of exact namespaces and identifiers, placeholders were accepted of the format "PH:placeholder" (see term and function level for examples). If a full statement is correct but BEL terms (representing entities) are expressed as placeholders instead of namespaces and identifiers, only a FN (false negative) but no FP (false positive) is counted. Similarly, the relationship type 'associate' (short form '--') could be used if the relationship type and/or the direction is unknown. The benefit in allowing such placeholders is to permit participants to include possibly relevant partial statements without suffering a penalization in their precision scores.

BEL statements can be submitted on different levels as full BEL statements or as fragments of full BEL statements. A submitted full BEL statement is automatically cut into its fragments to ensure evaluations on lower levels. Moreover, submissions on different levels were feasible too. A maximum number of three submissions were allowed in task 1. An example of a candidate evaluation is shown in Figure 4.

Sentence:

Sent.-Id:10004582 PMID:15909112 In the present study, we found that transgenic mice overexpressing wild-type human APP gene (hAPP/+) displayed a much higher expression of FAS, one of the death receptor subfamily.

BEL statements in gold standard and prediction

Sent.-Id	Gold standard BEL statement	Prediction BEL statement
10004582	p(HGNC:APP) -> p(HGNC:FAS)	act(p(HGNC:APP)) -> bp(GOBP:"gene expression")
		act(p(HGNC:APP)) -> act(p(HGNC:FAS))

Sentence based evaluation

Sent.-Id	Class	TP	FP	FN	Recall	Precision	F-score
10004582	Term (T)	2	1	0	100.00	66.67	80.00
10004582	Function-Secondary (FS)	0	1	0	0	0	0
10004582	Function (F)	0	2	0	0	0	0
10004582	Relation-Secondary (RS)	1	0	0	100.00	100.00	100.00
10004582	Relation (R)	1	1	0	100.00	50.00	66.67
10004582	Statement (S)	0	2	1	0	0	0

Fig. 4. An example result page of a candidate evaluation. The example shows the candidate sentence. Also the gold and predicted statements are provided. The calculated evaluation scores are shown for all primary and secondary levels.

On term level, only the correctness of BEL terms is evaluated. BEL terms are built from entities, their namespaces and associated abundance or process functions. The evaluation of BEL terms includes the correctness of the discovered entities, the correctness of associated namespaces and their format as well as the correctness of the associated abundance/process function.

On function level the correctness of discovered function is evaluated. Functions are only accepted together with their argument BEL terms. On the function level the correctness of functions together with their arguments is evaluated – it is TP (true positive) if the function is associated with the correct BEL terms. A complex function is valid if at least one of its arguments is correct. On the secondary function level, the correctness of a function alone was measured, regardless of the correctness of their term-arguments but with the presence of a BEL terms or placeholder.

In the relationship-level evaluation, only the entities and the relationships are considered. In general, functions that are part of a BEL statement are not taken into account on this level. In the special case of the *complex()* function, one correct function argument being in a correct relationship is sufficient for a positive evaluation. At the relationship level there are yet again two levels of evaluation considered. For a full-score relationship, subject, object, as well as the relationship type must be correct. For the secondary relationship level, partial relationships, containing two correct units out of three (subject, object and relationship type), are considered fulfilled.
Finally, we evaluated how many BEL statements are entirely correct. Submission of fragments of BEL statements can score higher in other levels but will damage the full statement level. BEL statements containing placeholder insertions are ignored in the full relationship level and full statement level evaluation.

6 Results

6.1 Task 1: Given textual evidence for a BEL statement, generate the corresponding BEL statement.

Five teams contributed information extraction systems for task 1. Each team was permitted to provide up to 3 runs. Table 5a and b shows the results for the task in stage 1 where the teams had to provide their own term recognition. The results are color-coded in shades of green according to the values of F-score (F), the main evaluation criterion, and supplemented by the values for precision (P) and recall (R). The best results for each evaluation metrics are marked up in bold.

For the full statement level, the best system s3 [12] achieved 20% F-measure, which illustrates the difficulty of this highly structured prediction task. System s4 [11] and s5 [13] had a similar performance, although their results were quite different on other evaluation levels, e.g. the term level. Obviously, the performance on the function level does not correlate well with the performance of the full statement level. One of the reasons is the lack of functions in 39 statements out of 105 test set statements. Furthermore, high scores on the relation level do not necessarily correlate with high scores on the full statement level. On the secondary relation level where only two thirds of the relationship has to be correct, up to 72.7% F-score were achieved.

		Terms			Function			Function Second.		
Sys	**Run**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
s1	r1	32.4	38.0	28.3	11.8	26.3	7.6	36.6	**86.7**	23.2
s2	r1	53.2	50.5	56.3	13.4	11.2	16.7	26.0	22.7	30.4
	r2	53.9	49.4	59.3	13.9	11.2	18.2	26.5	22.5	32.1
	r3	56.2	52.6	60.3	13.6	11.5	16.7	23.7	20.3	28.6
s3	r1	34.0	**84.2**	21.3	8.6	**75.0**	4.6	10.0	75.0	5.4
	r2	33.8	81.0	21.3	8.5	60.0	4.6	13.1	80.0	7.1
	r3	33.8	81.0	21.3	8.2	42.9	4.6	16.1	83.3	8.9
s4	r1	45.0	67.8	33.7	2.7	12.5	1.5	9.5	42.9	5.4
	r2	53.6	67.9	44.3	2.7	12.5	1.5	9.5	42.9	5.4
	r3	62.6	64.2	**61.0**	0.0	0.0	0.0	0.0	0.0	0.0
s5	r1	**68.9**	82.0	59.3	32.1	27.8	**37.9**	**54.6**	50.8	**58.9**
	r2	62.5	83.3	50.0	**32.6**	30.7	34.9	53.2	54.7	51.8
ensemble		28.0	98.0	16.3	5.8	66.7	3.0	3.5	50.0	1.8

		Relation			Relation Second.			Statement		
Sys	**Run**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
s1	r1	1.3	1.2	1.5	23.3	20.6	26.7	0.9	0.8	1.0
s2	r1	7.2	8.3	6.4	58.7	58.0	59.4	4.5	5.2	4.0
	r2	8.9	9.5	8.4	59.5	55.6	63.9	6.4	6.8	5.9
	r3	9.0	9.7	8.4	63.2	60.0	66.8	7.0	7.6	6.4
s3	r1	25.1	60.4	15.8	41.4	91.5	26.7	**20.2**	**54.4**	12.4
	r2	24.8	57.1	15.8	40.9	87.1	26.7	19.9	51.0	12.4
	r3	24.6	55.2	15.8	40.9	87.1	26.7	19.8	49.0	12.4
s4	r1	26.4	39.6	19.8	56.7	82.9	43.1	19.7	31.2	14.4
	r2	26.3	34.4	21.3	62.3	78.8	51.5	19.5	26.7	**15.4**
	r3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
s5	r1	49.2	69.4	38.1	71.8	76.8	**67.3**	18.2	26.4	13.9
	r2	49.2	69.4	38.1	**72.7**	**92.4**	59.9	18.2	26.4	13.9
ensemble		24.1	93.3	13.9	32.8	95.2	19.8	20.2	88.5	11.4

Table 5. Evaluation of stage 1 of task 1 (prediction of BEL statements without gold standard entities)

		Terms			Function			Function Second.		
Sys	**Run**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
s1	r1	**96.0**	96.9	**95.0**	5.6	40.0	3.0	10.2	**100.0**	5.4
s2	r1	61.0	87.0	47.0	10.7	13.0	9.1	24.3	20.2	30.4
	r2	64.7	85.7	52.0	10.3	12.0	9.1	23.5	19.1	30.4
	r3	62.5	80.5	51.0	10.5	12.5	9.1	22.9	19.1	28.6
s3	r1	54.3	**97.4**	37.7	20.8	**72.7**	12.1	26.1	69.2	16.1
s4	r1	55.2	96.7	38.7	0.0	0.0	0.0	0.0	0.0	0.0
	r2	51.7	96.4	35.3	0.0	0.0	0.0	0.0	0.0	0.0
	r3	70.9	96.6	56.0	0.0	0.0	0.0	0.0	0.0	0.0
s5	r1	82.4	91.8	74.7	30.0	25.5	**36.4**	**56.5**	51.5	**62.5**
	r2	79.7	92.5	70.0	**30.5**	27.1	34.9	54.2	51.6	57.1
ensemble		64.6	97.3	48.3	8.5	60.0	4.6	10.0	75.0	5.4

		Relation			Relation Second.			Statements		
Sys	**Run**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
s1	r1	25.9	21.3	33.2	**86.4**	81.0	**92.6**	14.7	12.5	17.8
s2	r1	6.1	26.9	3.5	55.8	65.8	48.5	3.5	16.7	2.0
	r2	10.0	31.6	5.9	57.9	63.2	53.5	7.6	25.0	4.5
	r3	9.6	25.5	5.9	58.0	64.1	53.0	8.1	22.2	5.0
s3	r1	43.7	75.6	30.7	61.5	96.8	45.1	**35.2**	67.6	**23.8**
s4	r1	44.6	81.6	30.7	63.5	**100.0**	46.5	33.1	68.8	21.8
	r2	42.1	82.6	28.2	61.2	**100.0**	44.1	30.8	**69.0**	19.8
	r3	45.5	66.0	34.7	76.7	97.0	63.4	32.9	53.3	**23.8**
s5	r1	65.1	77.9	55.9	82.4	87.7	77.7	25.6	32.1	21.3
	r2	65.1	77.9	55.9	83.4	94.4	74.8	25.6	32.1	21.3
ensemble		51.4	80.9	37.6	70.2	95.7	55.5	39.0	72.0	26.7

Table 6. Evaluation of stage 2 of task 1 (prediction of BEL statements with gold standard entities)

In a final step, we explored whether the performance can be enhanced through ensemble solutions. Considering all submitted statements of the five teams, the recall reaches 32.2% (best individual system run achieves 15.4%) but the precision drops to 9.2%. As result, the F-measure of 14.3% is substantially lower compared to the best individual system (data not shown). An ensemble system that considers all statements predicted by at least 2 different systems performs on F-measure level on par with the best individual system (c.f. Table 5). However, precision was gained at the expense of lower recall. Overall, the upper limit on recall is quite low: for 62 sentences (59%), no participating system could find any correct BEL statement. On the level of relations, 42 sentences (40%) had no true positive.

Table 6 shows the results for stage 2 of task 1 where the gold standard terms of the test set were made available to the teams. Most systems strongly benefit and improve on the level of full statements. These results prove again that high-quality relation extraction crucially depends on high-quality term recognition. With this setting, system s3 can compensate its rather low recall on the level of terms and can reach the best F-measure of 35.2% on the level of full statements. In this stage, considering all

statements predicted by at least 2 different systems outperforms the best individual system by almost 4%. The number of sentences where no system predicts any correct BEL statement dropped from 62 to 44 sentences (42%). On the level of relations, 19 sentences still had no true positive.

6.2 Task 2: Given a BEL statement, provide at most 10 additional evidence sentences.

For this task only one team participated [14]. The correctness of the provided evidence sentences (up to 10 sentences for each BEL statement) was evaluated manually and rated on three different levels of strictness:

1. Full: Relationship is fully expressed in the sentence.
2. Relaxed: Relationship can be extracted from the sentence if context sentences or biological background knowledge are taken into account.
3. Context: The sentence provides a valid context for the relationship, the entities are described by the sentence but the correct relation may not be expressed.

The system provided 806 evidence sentences for 96 BEL statements (mean 8.3 sentences per statement with a standard deviation 3.0). For 72 BEL statements, there was at least one entirely correct evidence sentence, for 78 statements at least one sentence meeting the relaxed evaluation conditions, and for 81 a sentence meeting the contextual conditions. Table 7 shows the detailed numbers for TP, FP and the resulting precision at the micro level. A bit more than one third of all sentences fully expressed the desired relationship. In order to assess the ranking quality of the system, we computed the mean average precision (MAP) and compared it with three alternative ranking scenarios:

- **Worst**: All TP are ranked after all false positives.
- **Random**: We randomly reordered the results 2000 times and computed the average MAP for all these variants.
- **Best**: All TP are ranked before all FP.

Table 7 shows that the system performs consistently better than random ranking but there is also capacity for improvement.

Criterion	TP	FP	Precision	MAP	Worst	Random	Best
Full	316	490	39.2%	49.0%	31.7%	46.5%	74.2%
Relaxed	429	377	53.2%	62.1%	45.9%	58.4%	80.4%
Context	496	310	61.5%	68.9%	55.2%	65.7%	83.5%

Table 7. Evaluation results of task 2 including mean average precision (MAP)

7 Acknowledgment

We would like to thank Natalie Catlett and William Hayes from Selventa for providing their time and expertise in helping us understand BEL. We would like to thank the BEL curators Alpha Tom Kodamullil and Reagon Karki for their evaluation support for task 2. We acknowledge support of our research from Philip Morris International.

REFERENCES

1. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 2003; 19:524–531
2. Demir E, Cary MP, Paley S, et al. The BioPAX community standard for pathway data sharing. Nat. Biotechnol. 2010; 28:935–942
3. Slater T, Song D. Saved by the BEL: ringing in a common language for the life sciences. 2012;
4. De León H, Boué S, Schlage WK, et al. A vascular biology network model focused on inflammatory processes to investigate atherogenesis and plaque instability. J. Transl. Med. 2014; 12:185
5. Schlage WK, Westra JW, Gebel S, et al. A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue. BMC Syst. Biol. 2011; 5:168
6. Gebel S, Lichtner RB, Frushour B, et al. Construction of a computable network model for DNA damage, autophagy, cell death, and senescence. Bioinform. Biol. Insights 2013; 7:97–117
7. Westra JW, Schlage WK, Frushour BP, et al. Construction of a computable cell proliferation network focused on non-diseased lung cells. BMC Syst. Biol. 2011; 5:105
8. Boue S, Fields B, Hoeng J, et al. Enhancement of COPD biological networks using a web-based collaboration interface. F1000Research 2015; 4:32
9. Boué S, Talikka M, Westra JW, et al. Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. Database (Oxford). 2015; 2015:bav030–
10. Liu W, Islamaj Doğan R, Kwon D, et al. BioC implementations in Go, Perl, Python and Ruby. Database (Oxford). 2014; 2014
11. Po-Ting Lai, Yu-Yan Lo, Ming-Siang Huang, Yu-Cheng Hsiao and Richard Tzong-Han Tsai. NCU-IISR System for BioCreative BEL Task 1. In: Proceedings of the fifth BioCreative challenge evaluatio workshop. Sevilla, Spain. 2015
12. Miji Choi, Haibin Liu, William Baumgartner, Justin Zobel and Karin Verspoor. Integrating Coreference Resolution for BEL Statement Generation. In: Proceedings of the fifth BioCreative challenge evaluatio workshop. Sevilla, Spain. 2015
13. Ravikumar Komandur Elayavilli, Majid Rastegar-Mojarad and Hongfang Liu. Adapting a rule-based relation extraction system for BioCreative V BEL task. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Sevilla, Spain. 2015
14. Majid Rastegar-Mojarad, Ravikumar Komandur Elayavilli and Hongfang Liu. Retrieving evidence sentences for BEL statements.. In: Proceedings of the fifth BioCreative challenge evaluation workshop. Sevilla, Spain. 2015