Evaluating large-scale text mining applications beyond the traditional numeric performance measures

Sofie Van Landeghem[1,2], **Suwisa Kaewphan**[3,4], **Filip Ginter**[3], **Yves Van de Peer**[1,2]
1. Dept. of Plant Systems Biology, VIB, Belgium
2. Dept. of Plant Biotechnology and Bioinformatics, Ghent University, Belgium
3. Dept. of Information Technology, University of Turku, Finland
4. Turku Centre for Computer Science (TUCS), Turku, Finland
solan@psb.ugent.be, sukaew@utu.fi
ginter@cs.utu.fi, yvpee@psb.ugent.be

Abstract

Text mining methods for the biomedical domain have matured substantially and are currently being applied on a large scale to support a variety of applications in systems biology, pathway curation, data integration and gene summarization. Community-wide challenges in the BioNLP research field provide gold-standard datasets and rigorous evaluation criteria, allowing for a meaningful comparison between techniques as well as measuring progress within the field. However, such evaluations are typically conducted on relatively small training and test datasets. On a larger scale, systematic erratic behaviour may occur that severely influences hundreds of thousands of predictions. In this work, we perform a critical assessment of a large-scale text mining resource, identifying systematic errors and determining their underlying causes through semi-automated analyses and manual evaluations[1].

1 Introduction

The development and adaptation of natural language processing (NLP) techniques for the biomedical domain are of crucial importance to manage the abundance of available literature. The inherent ambiguity of gene names and complexity of biomolecular interactions present an intriguing challenge both for BioNLP researchers as well as their targeted audience of biologists, geneticists and bioinformaticians. Stimulating such research, various community-wide challenges have been organised and received international participation.

The BioCreative (BC) challenge (Hirschman et al., 2005; Krallinger et al., 2008; Leitner et al., 2010; Arighi et al., 2011) touches upon a variety of extraction targets. The identification of gene and protein mentions ('named entity recognition') is a central task and a prerequisite for any follow-up work in BioNLP. Linking these mentions to their respective gene database identifiers, 'gene normalization', is a crucial step to allow for integration of textual information with authoritative databases and experimental results. Other BC tasks are engaged in finding functional and physical relations between gene products, including Gene Ontology annotations and protein-protein interactions.

Focusing more specifically on the molecular interactions between genes and proteins, the BioNLP Shared Task on Event Extraction (Kim et al., 2009; Kim et al., 2011b; Nedellec and others, 2013) covers a number of detailed molecular event types, including binding and transcription, regulatory control and post-translational modifications. Additionally, separate tracks involve specific applications of event extraction, including infectious diseases, bacterial biotopes and cancer genetics.

Performance of the participants in each of these challenges is measured using numeric metrics such as precision, recall, F-measure, slot error rate, MAP and TAP scores. While such rigurous evaluations allow for a meaningful comparison between different systems, it is often difficult to translate these numeric values into a measurement of practical utility when applied on a large scale. Additionally, infrequent but consistent errors are often not identified through small-scale evaluations, though they may result in hundreds of thousands of wrongly predicted interactions on a larger scale. In this work, we perform an in-depth study of an open-source state-of-the-art event extraction system which was previously applied to the whole of PubMed. Moving beyond the traditional numeric evaluations, we identify a num-

[1]The supplementary data of this study is freely available from http://bioinformatics.psb.ugent.be/supplementary_data/solan/bionlp13/

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 63–71,
Sofia, Bulgaria, August 4-9 2013.

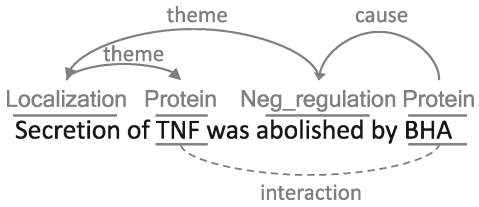

Figure 1: Example event and relation representations, depicted in solid and dotted lines respectively. Picture by Kim et al. (2011a).

ber of systematic errors in the large-scale data, analyse their underlying causes, and design post-processing rules to resolve these errors. We believe these findings to be highly relevant for any practical large-scale implementation of BioNLP techniques, as the presence of obvious mistakes in a text mining resource might undermine the credibility of text mining techniques in general.

2 Data and methods

In this section, we first describe the data and methods used in previous work for the construction of the large-scale text mining resource that is the topic of our error analyses (Section 3).

2.1 Event extraction

Event extraction has become a widely studied topic within the field of BioNLP following the first Shared Task (ST) in 2009. The ST'09 introduced the event formalism as a more detailed representation of the common binary relation annotation (Figure 1). Each event occurrence consists of an event trigger; i.e. one or more consecutive tokens that are linked to a specific event type. While the ST'09 included only 9 event types, among which 3 regulatory event types, the ST'11 further broadened the coverage of event extraction to post-translational modifications and epigenetics (EPI).

To compose a fully correct event, an event trigger needs to be connected to its correct arguments. Within the ST, these arguments are selected from a set of gold-standard gene and gene product annotations (GGPs). The ST guidelines determine an unambiguous formalism to which correct events must adhere: most event types only take one *theme* argument, while *Binding* events can be connected to more than one *theme*. *Regulation* events further have an optional *cause* slot (Figure 1). Connecting the correct arguments to the correct trigger words is denoted as 'edge detection'.

To perform event extraction, we rely on the publicly available Turku Event Extraction System (TEES) (Björne et al., 2012), which was originally developed for the ST'09. The TEES modules for trigger and edge detection are based upon supervised learning principles, employing support vector machines (SVMs) for multi-label classification. TEES has been shown to obtain state-of-the-art performance when measured on the gold-standard datasets of the Shared Tasks of 2009, 2011 and 2013.

2.2 Large-scale processing

Previously, the whole of PubMed has been analysed using a large-scale event extraction pipeline composed of the BANNER named entity recognizer, the McClosky-Charniak parser, and the Turku Event Extraction System (Björne et al., 2010). BANNER identifies gene and protein symbols in text through a machine learning approach based on conditional random fields (Leaman and Gonzalez, 2008). While the resulting large-scale text mining resource EVEX was focused only on abstracts and ST'09 event types (Van Landeghem et al., 2011), it has matured substantially during the past few years and now includes ST'11 EPI event types, full-text processing and gene normalization (Van Landeghem et al., 2013a). In this work, we use the version of EVEX as publicly available on 16 March 2013, containing 40 million event occurrences among 122 thousand gene and protein symbols in 22 million PubMed abstracts and 460 thousand PubMed Central full-text articles. Each event occurrence is linked to a normalized confidence value, automatically derived from the original TEES SVM classification step and the distance to the hyperplane of each prediction.

While this study focuses on the EVEX resource as primary dataset, the findings are also highly relevant for other large-scale text mining resources, especially those based on supervised learning, such as the BioContext (Gerner et al., 2012).

2.3 Cross-domain evaluation

Recently, a plant-specific, application-oriented assessment of the EVEX text mining resource has been conducted by manually evaluating 1,800 event occurrences (Van Landeghem et al., 2013b). In that study, it was established that the general performance rates as measured previously on the ST, are transferrable also to other domains and organisms. Specifically, the 58.5% TEES precision

Event type	Five most frequent trigger words				
Binding	binding	interaction	associated	bind	association
Gene expression	expression	expressed	production	expressing	levels
Localization	secretion	release	localization	secreted	localized
Protein catabolism	degradation	degraded	cleavage	proteolysis	degrade
Transcription	transcription	expression	levels	transcribed	detected
Acetylation	acetylation	acetylated	deacetylation	hyperacetylation	*activation*
Glycosylation	glycosylated	glycosylation	attached	N-linked	*absence*
Hydroxylation	hydroxylation	hydroxylated	hydroxylate	beta-hydroxylation	hydroxylations
Methylation	*radiation*	methylation	methylated	*diffractometer*	trimethylation
DNA methylation	methylation	hypermethylation	methylated	hypermethylated	unmethylated
Phosphorylation	phosphorylation	phosphorylated	dephosphorylation	phosphorylates	phosphorylate
Ubiquitination	ubiquitination	ubiquitinated	ubiquitylation	*ubiquitous*	polyubiquitination
Regulation	effect	regulation	effects	regulated	control
Positive regulation	increased	activation	increase	induced	induction
Negative regulation	reduced	inhibition	decreased	inhibited	inhibitor
Catalysis	mediated	dependent	mediates	removes	induced

Table 1: The top-5 most frequently tagged trigger words per event type in EVEX. The first 5 rows represent fundamental event types, the next 7 post-translational modifications (PTMs), and the last 4 rows are regulatory event types. In this analysis, the PTMs and their reverse types are pooled together. Trigger words that refer to systematic errors are in italic and are discussed further in the text.

rate measured in the ST'09, with the literature data concerning human blood cell transcription factors, corresponded with a 58.6% precision rate for the plant-specific evaluation dataset ('PLEV'). This encouraging result supports the general applicability of large-scale text mining methods trained on relatively small corpora. The findings of this previous study and the resulting data are further interpreted and analysed in more detail in this study.

3 Results

While the text mining pipeline underlying the EVEX resource has been shown to produce state-of-the-art results which are transferrable across domains and organisms, it is conceivable that the mere scale of the resource allows the accumulation of systematic errors. In this section, we perform several targeted semi-automated evaluations to identify, explain and resolve such cases. It is important to note that our main focus is on improving the precision rate of the resource, rather than the recall, aiming to increase the credibility of large-scale text mining resources in general.

3.1 Most common triggers

The trigger detection algorithm of the TEES software is based upon SVM classifiers (Section 2.1), and has been shown to outperform dictionary-based approaches (Kim et al., 2009; Kim et al., 2011c). To investigate its performance in a large-scale application, we first analyse the most frequent trigger words of each event type in EVEX (Table 1). We notice the presence of different inflections of the same word as well as related verbs and nouns, such as 'inhibition', 'inhibited' and 'inhibitor'. The trigger recognition module successfully uses character bigrams and trigrams in its SVM classification algorithm to allow for the identification of such related concepts, even when some of these trigger words were not encountered in the training phase (Björne et al., 2009).

However, occasionally this approach results in confusion between words with small edit distances, such as the trigger word 'ubiquitous' for *Ubiquitination* events. Similarly, the *Acetylation* trigger 'activation' is found within the context of a correct event structure in most cases, but should actually be of the type *Positive regulation*. The implementation of custom post-processing rules to automatically detect and resolve these specific cases would ultimately deal with more than 6,000 false-positive event predictions.

Further, the trigger 'radiation' seems to occur frequently for a *Methylation* event, of which 82% of the instances can be identified in the 'Experimental' subsection of the article. The majority of these articles relate to protein crystallography, and that subsection describes the data from the experimental set-up. Within such sections, phrases like 'Mo Kalpha radiation' are wrongly tagged as *Methylation* events. Similarly, many false-positive *Methylation* predictions refer to the trigger word 'diffractometer'. Removing these instances from the resource would result in the deletion of more

Trigger word s	Most frequent type t_2	Count	Frequency	Infrequent type t_1	Count	Frequency
acetylation	Acetylation	40,291	0.298383	Binding	1,332	0.000216
				Phosphorylation	1,050	0.001045
				Gene expression	969	0.000093
				Localization	1,045	0.000579
secretion	Localization	376,976	0.208888	Acetylation	243	0.001800
glycosylation	Glycosylation	24,226	0.141052	Phosphorylation	389	0.000387
				Gene expression	214	0.000020
phosphorylation	Phosphorylation	589,681	0.586772	Binding	454	0.000074
				DNA methylation	225	0.001297
ubiquitylation	Ubiquitination	4961	0.055976	Binding	128	0.000021
hypermethylation	Methylation	19,501	0.112434	Phosphorylation	365	0.000363
cleavage	Protein catabolism	20,552	0.073728	Gene expression	2,451	0.000234
				Binding	3,011	0.000489
decreased	Negative regulation	374,859	0.062372	Positive regulation	1,721	0.000173
				Binding	855	0.000139
				Gene expression	2,928	0.000280
reduced	Negative regulation	442,400	0.073610	Positive regulation	1,091	0.000110
reduction	Negative regulation	164,736	0.027410	Positive regulation	389	0.000039
absence	Negative regulation	65,180	0.010845	Positive regulation	226	0.000071

Table 2: Examples of trigger words that correspond to the type which has the highest relative frequency (left), but are also found with much lower frequencies in other types (right). The instances corresponding to the right-most column can thus be interpreted as wrong predictions. The full list is available as a machine readible translation table in the supplementary data.

than 82,000 false-positive event predictions.

Finally, we notice that the trigger word 'absence' for *Glycosylation* usually refers to a *Negative regulation*. Similarly, some words appear as most frequent for more than one event type, such as 'levels' (*Gene expression* and *Transcription*). This type of error in trigger type disambiguation is analysed in more detail in the next section.

3.2 Event type disambiguation

While previous work has focused on the disambiguation of event types on a small, gold-standard dataset (Martinez and Baldwin, 2011), the richness of a large-scale text mining resource provides additional opportunities to detect plausible errors. To exploit this large-scale information, we analyse all EVEX trigger words and their corresponding event types, summarizing their raw event occurrence counts as $Occ(t, s)$ where t denotes the trigger type and s the trigger string. As some event types are more abundantly described in literature, we normalize these counts to frequencies ($Freq(t, s)$) depending on the total number of event occurrences per type ($Tot(t)$):

$$Freq(t, s) = \frac{Occ(t, s)}{Tot(t)}$$

with

$$Tot(t) = \sum_{i=1}^{n} Occ(t, s_i)$$

and n the number of different triggers for event type t. We then compare all trigger words and their relative frequencies across different event types.

First, we inspect those cases where a trigger word appears with comparable frequencies for two event types t_1 and t_2:

$$Freq(t_1, s) \leq Freq(t_2, s) \leq 10 \times Freq(t_1, s) \quad (1)$$

A first broad category of these cases are trigger words that refer to both regulatory and non-regulatory events at the same time, such as 'overexpression' (*Gene expression* and *Positive regulation*), or 'ubiquitinates' (*Ubiquitination* and *Catalysis*). The majority of these cases are perfectly valid and are in fact modeled explicitly by the TEES software (Björne et al., 2009).

Further, we find that two broad groups of non-regulatory event types are semantically similar and share common trigger words: *Methylation* and *DNA methylation* (e.g. 'methylation', 'unmethylated', 'hypomethylation'), as well as *Gene expression* and *Transcription* ('expression', 'synthesis', 'levels'), with occasional overlap also with *Localization* ('abundance', 'found'). Similarly, trigger words are often shared among the four regulatory event types ('dependent', 'role', 'regulate'), as the exact type may depend on the broader context within the sentence.

While the previous findings do not necessar-

Curated event type	Predicted event type		
	Localization	**Transcription**	**Expression**
Localization	15	0	3
Transcription	0	12	1
Expression	0	2	12
No event	0	2	3
Total	15	16	19

Table 3: Targeted evaluation of 50 mixed events of type *Localization*, *Transcription* and *Gene expression*. The curated event type is compared to the original (hidden) predicted type.

ily refer to wrong predictions, we also notice the usage of punctuation marks as trigger words for various event types. This option was specifically provided in the TEES trigger detection algorithm as the ST'09 training data contains *Binding* instances with '-' as trigger word. However, these punctuation triggers are found to be largely false positives in the PubMed-scale event dataset. Removing them in an additional post-processing step would result in the filtering of more than 130,000 event occurrences, of which the largest part is expected to be incorrect predictions. Similarly, we can easily remove 25,000 events that are related to trigger words that are numeric values.

In a second step, we analyse those cases where

$$k \times Freq(t_1, s) \leq Freq(t_2, s). \qquad (2)$$

When this condition holds, it can be hypothesized that trigger predictions of the word s as type t_1 are false positives and should have instead been of type t_2. Automatically generating such lists from the data, we have experimentally determined an optimal value of k = 100 that represents a reasonable trade-off between the amount of false positives that can be identified and the manual work needed for this.

From the resulting list, we can easily identify a number of such cases that are clearly incorrect (Table 2, right column). Specifically, a large number of *Positive regulation* events actually refer to *Negative regulation*, providing an explanation of the lower precision rate of *Positive regulation* predictions in the previous PLEV evaluation (Van Landeghem et al., 2013b). This semi-automated detection procedure can ultimately result in the correction of more than 242,000 events.

The remaining cases for which condition (2) holds are more ambiguous and can not be automatically corrected. However, these cases are more likely to be incorrect and their confidence values could thus be automatically decreased depending on the ratio between $Freq(t_1, s)$ and $Freq(t_2, s)$. A general exception to this rule is formed by the broad groups of semantically similar events, such as *Transcription-Gene expression-Localization*, which we analyse in more detail in the next section.

3.3 Gene expression, Transcription and Localization

Transcription is a sub-process of *Gene expression*, with both event types relating to protein production. However, the distinction between the two in text may not always be straightforward. Additionally, the ST training data for *Transcription* events is significantly smaller than for *Gene expression* events, which may be the reason why not only the TEES performance, but also those of other systems, is considerably lower for *Transcription* than for *Gene expression* (Kim et al., 2011c). Further, cell-type specific gene expression should be captured by additional *site* arguments connected to a *Localization* event, which represents the presence or a change in the location of a protein.

To gain a deeper insight into the interplay between these three different event types, we have performed a manual curation of 50 event occurrences, sampled at random from the *Gene expression*, *Transcription* and *Localization* events available in EVEX. For each event, the trigger word and the corresponding sentence was extracted, but the predicted event type was hidden. An expert annotator subsequently decided on the correct event type of the trigger. Within this evaluation we followed the ST guidelines to only annotate *Gene expression* when there is no evidence for the more detailed *Transcription* type.

Table 3 shows the results. All 15 predicted *Localization* triggers are recorded to be correct. From the 16 predicted *Transcription* events, two involve incorrect event triggers, and two other events refer to the more general *Gene expression* type (75% overall precision). Likewise, only one *Gene expression* event should be of the more spe-

	Curated event type	Error type	Instances	(%)
1	Single-argument Binding	No error	5	10%
2	Single-argument Binding	Edge detection error	0	0%
3	Multiple-argument Binding	Edge detection error	4	8%
4	Single-argument Binding	Entity recognition error	1	2%
5	Multiple-argument Binding	Entity recognition error	19	38%
6	Other	Trigger detection error	21	42%

Table 4: Targeted evaluation of 50 single-argument *Binding* event triggers. Row 1: Fully correct event. Row 2: The correct argument was annotated but not linked. Row 3: At least one correct multiple-argument *Binding* event could have been extracted using the annotated entities in the sentence. Row 4: The correct argument was not annotated. Row 5: No event could be extracted due to missing argument annotations. Row 6: The trigger did not refer to a *Binding* event.

Unannotated entity type	Entity occurrence count	Examples
GGP	10	SPF30, spinal muscular atrophy gene
Generic GGP	9	primary antibodies, peptides, RNA
Chemical compound	10	Ca(2+), iron, manganese(II)

Table 5: Manual inspection of the textual entity types for those *Binding* events where a relevant *theme* argument was not annotated in the entity recognition step.

cific *Transcription* type, three instances should be *Localization*, and three more are considered not to be correct events at all (63% overall precision). In general, we remark that the predicted event type largely corresponds to the curated type (78% of all predictions and 87% of all otherwise correct events).

3.4 Binding

Moving beyond the event type specification as determined by the ST guidelines, the previous PLEV analysis (Section 2.3) has established a remarkable difference between single-argument and multiple-argument *Binding*. In contrast to the regular ST evaluations, this work considered single- and multiple-argument *Binding* as two separate event types, resulting in a precision rate of 93% for multiple-argument *Binding* triggers and only 8% precision for single-argument *Binding* triggers.

As the PLEV study only focused on textual network data, single-argument *Bindings* were not analysed further. In this work however, we further investigate this performance discrepancy and perform an in-depth manual evaluation to try and detect the main causes of this systematic error.

Several hypotheses can be postulated to explain the low precision rate of single-argument *Binding* events. Firstly, a false negative instance of the entity recognition module might result in the absence of annotation for a relevant second interaction partner. Another plausible explanation is an error by the edge detection module of the event extraction mechanism, which would occasionally decide to produce one or several single-argument *Binding* events rather than one multiple-argument *Binding*, even when all involved entities are correctly annotated. Finally, it is conceivable that predicted single-argument triggers simply do not refer to *Binding* events, i.e. they contain false positive predictions of the trigger detection module of the event extraction system.

In some cases, one trigger leads to many different *Binding* events, such as the trigger 'bind' in the sentence "*Sir3 and Sir4 bind preferentially to deacetylated tails of histones H3 and H4"*. In these cases, error types may accumulate: some events could be missed due to unannotated entities, while others may be due to errors in the edge detection step. However, multiple events with the same trigger word are often represented by very similar feature vectors in the classification step, and consequently have almost identical final confidence values. For this reason, we summarize the error as 'Edge detection error' as soon as one pair of entities was correctly annotated but not linked, and as 'Entity recognition error' otherwise.

Table 4 summarizes the results of a curation effort of 50 event triggers linked to a single-argument *Binding* event in EVEX. We notice that in fact, 46% should have been multiple-argument *Binding* events. The main underlying reason for the prediction of an incorrect single-argument *Binding* event, when it should have been a multiple-argument one, is apparently caused by

	Curated event type	Error type	Instances	(%)
1	Phosphorylation	No error	34	68%
2	Phosphorylation	Edge detection error	4	8%
3	Invalid Phosphorylation	Edge detection error	2	4%
4	Phosphorylation	Edge directionality detection error	4	8%
5	Invalid Phosphorylation	Edge directionality detection error	1	2%
6	Phosphorylation	Entity recognition error	3	6%
7	Other	Trigger detection error	2	4%

Table 6: Targeted evaluation of 50 *Phosphorylation* event triggers and their *theme* arguments. Row 1: Fully correct event. Row 2: The correct argument was annotated but not linked. Row 3: An argument was linked but should not have been. Row 4: A causal argument was wrongly annotated as the *theme* argument. Row 5: A causal argument was wrongly annotated as the *theme* argument. Row 6: The correct argument was not annotated. Row 7: The trigger did not refer to a *Phosphorylation* event.

an entity recognition error (19/23 or 83%), while an edge detection error is much less frequent (17%). When we examine these entity recognition errors in more detail, we find that 10 relevant entities are true GGPs in the sense of the Shared Task annotation. However, 9 entities refer to generic GGPs, and 10 instances relate to chemical compounds (Table 5). As these type of entities can not be unambiguously normalized to unique gene identifiers, they fall out-of-scope of the original ST challenge. However, we feel this practice introduces an artificial bias on the classifier and the evaluation. Additionally, this information can prove to be of value within a large-scale text mining resource geared towards practical applications and explorative browsing of textual information.

Finally, we notice that a remarkable 42% of all predicted events contain trigger detection errors. Analysing this subclass in more detail, we found that 5 cases are invalid event triggers, 6 cases refer to other event types such as *Localization* and *Gene expression*, and 10 more cases were considered to be out-of-scope of the ST challenge, such as a factor-disease association.

3.5 Phosphorylation

Within the PLEV evaluation (Section 2.3), it became apparent that *Phosphorylation* is easy to recognise from the sentence (98%) but the full correct event has a much lower precision rate (65%). As we have seen in the previous section, even when a trigger word is correctly predicted, errors may still be generated by the edge detection or entity recognition step. For instance, we might hypothesize that the main underlying reason for the reduced final performance is an error by the entity recognition step, forcing the edge detection mechanism to link an incorrect *theme* due to lack of other options. Other plausible explanations involve genuine errors by the edge detection algorithm when the correct argument is annotated, as well as problems with the identification of causality. As the TEES version applied in this work was developed for the Shared Task 2009 and 2011, it does not predict causal arguments for a *Phosphorylation* event directly, but instead adds *Regulation* events on top of the *Phosphorylations*. Occasionally, we have noticed that the *theme* of a *Phosphorylation* event should in fact have been the *cause* of the embedding *Regulation* association, resulting in a wrongly directed causal relationship.

To investigate these possibilities, we have manually inspected 50 *Phosphorylation* events picked at random from the EVEX resource. Table 6 summarizes the results of this effort. Only two events are found not to be *Phosphorylation* events: one is in fact a *Gene expression* mention, the other involves an incorrect trigger. Additionally, three more events can semantically be regarded as *Phosphorylations*, but do not follow the ST specifications ('Invalid Phosphorylation'), for instance because they only mention causal arguments ('an inhibition of Ca^{2+}/calmodulin-dependent protein phosphorylation'). Among the 45 cases which correctly refer to the *Phosphorylation* type, 34 events are fully correct (68% of the total). Four cases are wrongly extracted by misinterpreting the causal relationship ('Edge directionality detection error') and four more instances refer to genuine mistakes of the edge detection algorithm. Only three other cases can be attributed to a missing entity annotation. In contrast to the previous findings on single-argument *Bindings*, we thus establish that the incorrect *Phosphorylation* events are mainly caused by errors in the edge detection mechanism, which either picks the wrong *theme*

from the set of annotated GGPs, or misinterprets the causality direction.

4 Discussion and conclusion

We have performed several semi-automated evaluations and targeted manual curation experiments, identifying and explaining systematic errors in a large-scale event dataset. As a first observation, we notice that a few frequent trigger words are almost always associated to incorrect event predictions, such as the trigger words 'ubiquitous' and 'radiation', or a punctuation symbol. These cases were identified through a large-scale automatic analysis in combination with a limited manual evaluation effort. The results are distributed as a blacklist of event triggers for the implementation or filtering of future large-scale event predictions efforts.

Further, a semi-automated procedure has identified a list of likely incorrect predictions, by comparing the type-specific frequencies of trigger words across all event types. Manual inspection of the most frequent cases allowed us to determine a number of trigger words for which the event type can automatically be corrected. These results are also made publicly available.

Additionally, after removal of the most obvious and frequent errors, a fully automated script can automatically reduce the confidence scores of those event occurrences where the trigger words are found to be much more frequent for another event type. We have established that this procedure should disregard triggers identified within a few specific semantically similar clusters: *DNA methylation/Methylation*, *Regulation/Positive regulation/Negative regulation/Catalysis* and *Gene expression-Transcription/Localization*. An additional targeted evaluation of these last three types revealed that, despite their semantic overlap, the largest fraction of these predictions refers to the correct event type ($78 \pm 11.5\%$).

Finally, we note that trigger detection ($47 \pm 14.6\%$) and entity recognition errors ($44 \pm 14.6\%$) are the main causes of wrongly predicted *Binding* events. The latter causes the event extraction mechanism to artificially produce single-argument *Bindings* instead of multiple-argument *Bindings*. We believe this issue can be resolved by broadening the scope of the entity recognition module to generic GGPs and chemical compounds, and reapplying the TEES algorithm to these entities as if they were normal GGPs as defined in the ST formalism. In contrast, edge detection errors are much more frequently the cause of a wrongly predicted *Phosphorylation* event (statistically significant difference with $p < 0.05$), caused by wrongly identifying the thematic object or the causality of the event. To resolve this issue, we propose future annotation efforts to specifically annotate the protein adding the phosphate group to the target protein as a separate class than the regulation of such a phosphorylation process by other cellular machineries and components (Kim et al., 2013).

In conclusion, we have performed several statistical analyses and targeted manual evaluations on a large-scale event dataset. As a result, we were able to identify a set of rules to automatically delete or correct a number of false positive predictions (supplementary material at `http://bioinformatics.psb.ugent.be/supplementary_data/solan/bionlp13/`). When applying these rules to the winning submission of the recent ST'13 (GE subchallenge), which was based on the TEES classifier (Hakala et al., 2013), 3 false positive predictions could be identified and removed. Even though this procedure only marginally improves the classification results (50.97% to 50.99% F-score), we believe the cleaning procedure to be crucial specifically for the credibility of any large-scale text mining application. For example, applied on the EVEX resource, it would ultimately result in the removal of 242,000 instances and a corrected event type of 230,000 more cases (1.2% of all EVEX events in total). These corrections will be implemented as part of the next big EVEX release. Additionally, the confidence score of more than 120,000 ambiguous cases could be automatically decreased. Alternatively, these cases could be the target of a large-scale re-annotation, for instance using the brat annotation tool (Stenetorp et al., 2012). The resulting dataset could then serve as a new training set to enable active learning on top of existing event extraction approaches.

Acknowledgments

The authors thank Cindy Martens and the anonymous reviewers for a critical reading of the manuscript and constructive feedback. SVL thanks the Research Foundation Flanders (FWO) for funding her research.

References

Cecilia Arighi, Zhiyong Lu, Martin Krallinger, Kevin Cohen, J. Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy Wu. 2011. Overview of the BioCreative III workshop. *BMC Bioinformatics*, 12(Suppl 8):S1.

Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. 2009. Extracting complex biological events with rich graph-based feature sets. In *Proceedings of the BioNLP 2009 Workshop*, pages 10–18.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. 2010. Scaling up biomedical event extraction to the entire PubMed. In *Proceedings of the BioNLP 2010 Workshop*, pages 28–36.

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012. Generalizing biomedical event extraction. *BMC Bioinformatics*, 13(suppl. 8):S4.

Martin Gerner, Farzaneh Sarafraz, Casey M. Bergman, and Goran Nenadic. 2012. BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. *Bioinformatics*, 28(16):2154–2161.

Kai Hakala, Sofie Van Landeghem, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. EVEX in ST13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of the BioNLP Shared Task 2013 Workshop (in press)*.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. 2005. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. 2009. Overview of BioNLP'09 Shared Task on event extraction. In *Proceedings of the BioNLP 2009 Workshop*, pages 1–9.

Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Junichi Tsujii. 2011a. Extracting bio-molecular events from literature - the BioNLP'09 Shared Task. *Computational Intelligence*, 27(4):513–540.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. 2011b. Overview of BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011c. Overview of Genia event task in BioNLP Shared Task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, pages 7–15.

Jin-Dong Kim, Yue Wang, Yamamoto Yasunori, Sabine Bergler, Roser Morante, and Kevin Cohen. 2013. The Genia Event Extraction Shared Task, 2013 edition - overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop (in press)*.

Martin Krallinger, Alexander Morgan, Larry Smith, Florian Leitner, Lorraine Tanabe, John Wilbur, Lynette Hirschman, and Alfonso Valencia. 2008. Evaluation of text-mining systems for biology: overview of the second BioCreative community challenge. *Genome Biology*, 9(Suppl 2):S1.

Robert Leaman and Graciela Gonzalez. 2008. BANNER: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 652–663.

F. Leitner, S.A. Mardis, M. Krallinger, G. Cesareni, L.A. Hirschman, and A. Valencia. 2010. An overview of BioCreative II.5. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 7(3):385–399.

David Martinez and Timothy Baldwin. 2011. Word sense disambiguation for event trigger word detection in biomedicine. *BMC Bioinformatics*, 12(Suppl 2):S4.

Claire Nedellec et al. 2013. Overview of BioNLP Shared Task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop (in press)*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Sofie Van Landeghem, Filip Ginter, Yves Van de Peer, and Tapio Salakoski. 2011. EVEX: a PubMed-scale resource for homology-based generalization of text mining predictions. In *Proceedings of the BioNLP 2011 Workshop*, pages 28–37.

Sofie Van Landeghem, Jari Björne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013a. Large-scale event extraction from literature with multi-level gene normalization. *PLoS ONE*, 8(4):e55814.

Sofie Van Landeghem, Stefanie De Bodt, Zuzanna J. Drebert, Dirk Inz, and Yves Van de Peer. 2013b. The potential of text mining in data integration and network biology for plant research: A case study on arabidopsis. *The Plant Cell*, 25(3):794–807.