EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction

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

Abstract

During the past few years, several novel text mining algorithms have been developed in the context of the BioNLP Shared Tasks on Event Extraction. These algorithms typically aim at extracting biomolecular interactions from text by inspecting only the context of one sentence. However, when humans interpret biomolecular research articles, they usually build upon extensive background knowledge of their favorite genes and pathways. To make such world knowledge available to a text mining algorithm, it could first be applied to all available literature to subsequently make a more informed decision on which predictions are consistent with the current known data. In this paper, we introduce our participation in the latest Shared Task using the large-scale text mining resource EVEX which we previously implemented using state-of-the-art algorithms, and which was applied to the whole of PubMed and PubMed Central. We participated in the Genia Event Extraction (GE) and Gene Regulation Network (GRN) tasks, ranking first in the former and fifth in the latter.

1 Introduction

The main objective of our entry was to test the usability of the large-scale text mining resource EVEX to provide supporting information to an existing state-of-the-art event extraction system. In the GE task, EVEX is used to extract additional features for event extraction, capturing the occurrence of relevant events in other documents across PubMed and PubMed Central. In the GRN task, EVEX is the sole source of information, i.e. our entry consists of a modified subset of EVEX, rather than a new text mining system specifically trained for the task.

In the 2011 GE task, the majority of participating systems used features solely extracted from the immediate textual context of the event candidate, typically restricted to a single sentence (Kim et al., 2012; McClosky et al., 2012; Björne et al., 2012b; Vlachos and Craven, 2012). Several studies have subsequently incorporated coreference relations, capturing information also from surrounding sentences (Yoshikawa et al., 2011; Miwa et al., 2012). However, no prior work exists on extending the event context to the information extracted from other documents on a large scale. The motivation for this entry is thus to test whether a gain can be obtained by aggregating information across documents with mutually supporting evidence.

In the following sections, we first introduce EVEX as the underlying text mining resource, and then describe the methods developed specifically for the GRN and GE task entries. Finally, a detailed error analysis of the results offers insight into the performance of our systems and provides possible directions of future development.

2 EVEX

EVEX[1] is a text mining resource built on top of events extracted from all PubMed abstracts and PubMed Central Open-Access full-text documents (Van Landeghem et al., 2013a). The extraction was carried out using a combination of the BANNER named entity detector (Leaman and Gonzalez, 2008) and the TEES event extraction system as made publicly available subsequent to the last Shared Task (ST) of 2011 (Björne et al., 2012a). Specifically, this version of TEES was trained on the ST'11 GE data.

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

Proceedings of the BioNLP Shared Task 2013 Workshop, pages 26–34,
Sofia, Bulgaria, August 9 2013.

On top of the individual event occurrences, EVEX provides event generalizations, allowing the integration and summarization of knowledge across different articles (Van Landeghem et al., 2011). For instance, the canonicalization algorithm deals with small lexical variations by removing non-alphanumerical characters (e.g. 'Esr-1' to 'esr1'). The canonical generalization then groups those events together with the same event type and the same canonicalized arguments. Additionally, gene normalization data has recently been integrated within the EVEX resource, assigning taxonomic classification and database identifiers to gene mentions in text using the GenNorm system (Wei and Kao, 2011). Finally, the assignment of genes to homologous families allows a more coarse-grained generalization of the textual data. For each generalized event, a confidence score is automatically calculated based upon the original TEES classification procedure, with higher values representing more confident predictions.

Finally, the EVEX resource provides a network interpretation which transforms events into pairwise gene/protein relations to represent a typed, directed network. The primary advantage of such a network, as compared to the complex, recursive event structures, is that a network is more easily analysed and integrated with other external resources (Kaewphan et al., 2012; Van Landeghem et al., 2013b).

3 GRN Task

The Gene Regulatory Network subtask of the ST'13 aims at evaluating the ability of text mining systems to automatically compile a gene regulation network from the literature. The task is focused specifically on sporulation in *Bacillus subtilis*, a thoroughly studied process.

3.1 Challenge definition

The primary goal of our participation in this task was assessing the ability to reconstruct regulatory networks directly from the EVEX resource. Consequently, we have applied the EVEX data as it is publicly available. This decision has two major consequences. First, we have used the predicted BANNER entities rather than the gold-standard entity annotation, artificially rendering the challenge more difficult. Second, we did not adapt the EVEX events, which follow the ST'11 GE formalism, to the novel annotation scheme of the GRN challenge, but rather derived the network data directly from the EVEX interactions. For example, given these trigger annotations

```
T1 Protein 37 43 sigmaB
T2 Gene 54 58 katX
T3 Transcription 59 69 expression
```

a GE Transcription event looks like

```
E1 Transcription:T3 Theme:T2 Cause:T1
```

while the GRN annotation is given by

```
R1 Transcription Target:E1 Agent:T1
E1 Action_Target:T3 Target:T2
```

However, both formalisms can easily be translated into the required GRN network format:

```
sigB Interaction.Transcription katX
```

where 'sigB' is annotated as the *Gene identifier* of 'sigmaB'. These gene identifiers are provided in the gold-standard entity annotations. Note that in this context, "gene identifiers" are standardized gene symbols rather than numeric identifiers, and full gene normalization is thus not required.

EVEX type	GRN type
Binding	Binding
Regulation* of Transcription	Transcription
Regulation* of Gene expression	Transcription
Positive regulation of Any*	Activation
Negative regulation of Any*	Inhibition
Regulation of Any*	Regulation

Table 1: Conversion of EVEX event types to the GRN types. The table is traversed from top to bottom, and the first rule that matches is applied. Regulation* refers to any type of regulatory event, and Any* refers to any other non-regulatory event type.

3.2 From EVEX to GRN data

As a first step towards creating a gene regulatory network directly from EVEX, we have downloaded all pairwise relations of the canonical generalization (Section 2). For each such relation, we also obtain important meta-data, including the confidence value, the PubMed IDs in which a relation was found, whether or not those articles describe *Bacillus subtilis* research, and whether or not those articles are part of the GRN training or test set. In the most stringent setting, we could then limit the EVEX results only to those relations found in the articles of the GRN dataset (72 in training, 45 in the development set, 55 in the test set). Additionally, we could test whether performance can be improved by also adding all *Bacillus subtilis* articles (17,065 articles) or even

GRN event type	Possible target types	Possible agent types
Interaction.Binding	Protein	Gene
Interaction.Transcription	Protein, PolymeraseComplex	Gene, Operon
Interaction.Regulation Interaction.Activation Interaction.Inhibition	Protein, PolymeraseComplex	Gene, Operon, Protein, ProteinComplex

Table 2: Entity-type filtering of event predictions. Only those events for which the arguments (the target as well as the agent) have the correct entity types, are retained in the result set.

all EVEX articles in which at least one event was found (4,107,953 articles).

To match the canonicalized BANNER entities from EVEX to the standardized gene symbols required for the GRN challenge, we have constructed a mapping based on the GRN data. First, we have scanned all gold-standard entities and removed non-alphanumerical characters from the gene symbols as tagged in text. Next, these canonical forms were linked to the corresponding standardized gene symbols in the gold-standard annotations. From the EVEX data, we then only retained those relations that could be linked to two gene symbols occurring together in a sentence.

Finally, it was necessary to convert the original EVEX event types to the GRN relation types. This mapping is summarized in Table 1. Because EVEX Binding events are symmetrical and GRN Bindings are not, we add both possible directions to the result set. Note that some GRN types could not be mapped because they have no equivalent within the EVEX resource, such as the GRN type 'Requirement' or 'Promoter'.

3.3 Filtering the data

After converting the EVEX pairwise relations to the GRN network format, it is necessary to further process the set of predictions to obtain a coherent network. One additional filtering step concerns the entity types of the arguments of a specific event type. From the GRN data, we can retrieve a symbol-to-type mapping, recording whether a specific symbol referred to e.g. a gene, protein or operon in a certain article. After careful inspection of the GRN guidelines and the training data, we enforced the filtering rules as listed in Table 2. For example, this procedure successfully removes protein-protein interactions from the dataset, which are excluded according to the GRN guidelines. Even though these rules are occasionally more restrictive than the original GRN guidelines, their effectiveness to prune the data was confirmed on the training set.

Further, the GRN guidelines specify that a set of edges with the same *Agent* and *Target* should be resolved into a single edge, giving preference to a more specialized type, such as Transcription in favour of Regulation. Further, contradictory types between a specific entity pair (e.g. Inhibition and Activation) may occur simultaneously in the GRN data. For the EVEX data however, it is more beneficial to try and pick one single correct event type from the set of predictions, effectively reducing the false positive rate. To this end, the EVEX confidence values are used to determine the single most plausible candidate. Further analyses on the training data suggested that the best performance could be achieved when only retaining the 'Mechanism' edges (Transcription and Binding) in cases when no regulatory edge was found. Finally, we noted that the EVEX Binding events more often correspond to the GRN Transcription type, and they were thus systematically refactored as such (after entity-type filtering). We believe this shift in semantics is caused by the fact that a promoter binding is usually extracted as a binding event by the TEES classifier, while it can semantically be seen as a Transcription event, especially in those cases where the Theme is a protein name, and the Cause a gene symbol (Table 2).

3.4 Results

Table 3 lists the results of our method on the GRN training data, which was primarily used for tuning the parameters described in Section 3.3. The highest recall (42%) could be obtained when using all EVEX data, without restrictions on entity types and without restricting to *Bacillus subtilis* articles. As a result, this set of predictions may contain relations between homologs in related species which have the same name. While the relaxed F-score (41%) is quite high, the Slot Error Rate (SER) score (1.56) is unsatisfying, as SER scores should be below 1 for decent predictions.

When applying entity type restrictions to the prediction set, relaxed precision rises from 39%

Dataset	ETF	**SER**	**F**	**Rel. P**	**Rel. R**	**Rel. F**	**Rel. SER**
All EVEX data	no	1.56	8.86	39.29%	**41.98**%	40.59%	1.23
All EVEX data	yes	1.15	11.53	59.74%	35.11%	**44.23**%	0.89
B. subtilis PMIDs	yes	0.954	**20.81**	71.43%	22.90%	34.68%	**0.86**
GRN PMIDs	yes	**0.939**	17.39	**80.00**%	18.32%	29.81%	**0.86**

Table 3: Performance measurement of a few different system settings, applied on the training data. The SER score is the main evaluation criterion of the GRN challenge. The relaxed precision, recall, F and SER scores are produced by scoring the predictions regardless of the specific event types. ETF refers to entity type filtering.

to 60%, the relaxed F-score obtains a maximum score of 44%, and the SER score improves to 1.15. The SER score can further be improved when restricting the data to *Bacillus subtilis* articles (0.954). The optimal SER score is obtained by further limiting the prediction set to only those relations found in the articles from the GRN dataset (0.939), maximizing at the same time the relaxed precision rate (80%).

The final run which obtained the best SER score on the training data was subsequently applied on the GRN test data. It is important to note that the parameter selection of our system was not overfitted on the training data, as the SER score of our final submission on the test data is 0.92, i.e. higher than the best run on the training data.

Table 4 summarizes the official results of all participants to the GRN challenge. Interestingly, the TEES classifier has been modified to retrain itself on the GRN data and to produce event annotations in the GRN formalism (Björne and Salakoski, 2013), obtaining a final SER score of 0.86. It is remarkable that this score is only 0.06 points better than our system which needed no retraining, and which was based upon the original GE annotation format and predicted gene/protein symbols rather than gold-standard ones. Additionally, the events in EVEX have been produced by a version of TEES which was maximized on F-score rather than SER score, and these measurements are not mutually interchangeable (Table 3). We conclude that even though our GRN system obtained last place out of 5 participants, we believe that its relative close performance to the TEES submission demonstrates that large-scale text mining resources can be used for gene regulatory network construction without the need for retraining the text mining component.

3.5 Error analysis

To determine the underlying reasons of our relatively low recall rate, we have analysed the 117 false negative predictions of our final run on the training dataset. We found that 23% could be attributed to a missing or incompatible BANNER entity, 59% to a false negative TEES prediction, 15% to a wrong GRN event type and 3% to incorrectly mapping the gene symbol to the standardized GRN format. Analysing the 16 false positives in the same dataset, 25% could be attributed to an incorrectly predicted event structure, and 62.5% to a wrongly predicted event type. One case was correctly predicted but from a sentence outside the GRN data, and in one case a correctly predicted negation context was not taken into account. In conclusion, future work on the GRN conversion of TEES output should mainly focus on refining the event type prediction, while general performance could be enhanced by further improving the TEES classification system.

	SER	**Relaxed SER**
University of Ljubljana	0.73	0.64
K.U.Leuven	0.83	0.66
TEES-2.1	0.86	0.76
IRISA-TexMex	0.91	0.60
EVEX	0.92	0.81

Table 4: Official GRN performance rates.

4 GE Task

Our GE submission builds on top of the TEES 2.1 system[2] as available just prior to the ST'13 test period. First applying the unmodified TEES system, we subsequently re-ranked its output and enforced a cut-off threshold with the objective of removing false positives from the TEES output (Section 4.1). In the official evaluation, this step results in a minor 0.23pp increase of F-score compared to unprocessed TEES output (Table 5). This yields the *first rank* in the primary measure of the task with TEES ranking second.

The main motivation for the re-ranking ap-

[2] https://github.com/jbjorne/TEES/wiki/TEES-2.1

	P	R	F
EVEX	58.03	45.44	50.97
TEES-2.1	56.32	46.17	50.74
BioSEM	62.83	42.47	50.68
NCBI	61.72	40.53	48.93
DlutNLP	57.00	40.81	47.56

Table 5: Official precision, recall and F-score rates of the top-5 GE participants, in percentages.

proach was the ability to incorporate external information from EVEX to compare the TEES event predictions and identify the most reliable ones. Further, such a re-ranking approach leads to an independent component which is in no way bound to TEES as the underlying event extraction system. The component can be combined with any system with sufficient recall to justify output re-ranking.

4.1 Event re-ranking

The output of TEES is re-ranked using SVM^{rank}, a formulation of Support Vector Machines which is trained to optimize ranking, rather than classification (Joachims, 2006). It differs from the basic linear SVM classifier in the training phase, when a *query structure* is defined as a subset of instances which can be meaningfully compared among each other — in our case all events from a single sentence. During training, only instances within a single query are compared and the SVM does not aim to learn a global ranking across sentences and documents. We also experimented with polynomial and radial basis kernels, feature vector normalization and broadening the ranking query sets to whole sections or narrowing them to only events with shared triggers, but none of these settings were found to further enhance the performance.

The re-ranker assigns a numerical score to each event produced by TEES, and all events below a certain threshold score are removed. To set this threshold, a linear SVM regressor is applied with the SVM^{light} package (Joachims, 1999) to each sentence individually, i.e. we do not apply a data-wide, pre-set threshold. Unlike the re-ranker which receives features from a single event at a time, the regressor receives features describing the set of events in a single sentence.

Re-ranker features

Each event is described using a number of features, including the TEES prediction scores for triggers and arguments, the event structure, and the EVEX information about this as well as similar events. Events can be recursively nested, with the root event containing other events as its arguments. The root event is of particular importance as the top-most event. A number of features are thus dedicated specifically to this root event, while other features capture properties of the nested events.

Features derived from TEES confidence scores:

- TEES trigger detector confidence of the root event and its difference from the confidence of the negative class, i.e. the margin by which the event was predicted by TEES.
- Minimum and maximum argument confidences of the root event.
- Minimum and maximum argument confidences, including recursively nested events (if any).
- Minimum and maximum trigger confidences, including recursively nested events (if any).
- Difference between the minimum and maximum argument confidences compared to other events sharing the same trigger word.

Features describing the structure of the event:

- Event type of the root trigger.
- For each path in the event from the root to a leaf argument, the concatenation of event types along the path.
- For each path in the event from a leaf argument to another leaf argument, the concatenation of event types along the path.
- The event structure encoded in the bracketed notation with leaf (T)heme and (C)ause arguments replaced by a placeholder string, e.g.

  `Regulation(C:_, T:Acetylation(T:_)).`

Features describing other events in the same sentence:

- Event counts for each event type.
- Event counts for each unique event structure given by the bracketed structure notation.

All event counts extracted from EVEX are represented as their base-10 logarithm to compress the range and suppress differences in counts of very common events.

The following features are generated in two versions, one by grouping the events according to the EVEX *canonical* generalization and one for the *Entrez Gene* generalization (Section 2)[3].

[3]The generalizations based on gene families were evaluated as well, but did not result in a positive performance gain.

- All occurrences of the given event in EVEX.
- For each path from root to a leaf gene/protein, all occurrences of that exact path in EVEX.
- For each pair of genes/proteins in the event, all occurrences of that pair in the network interpretation of EVEX.
- For each pair of genes/proteins in the event, all occurrences of that pair with a different event type in the network interpretation of EVEX.

For each event, path, or pair under consideration, features are created for the base-10 logarithm of the count in EVEX and of the number of unique articles in which it was identified, as well as for the minimum, maximum, and average confidence values, discretized into six unique categories.

Regressor features

While the re-ranker features capture a single event at a time, the threshold regressor features aggregate information about events extracted within one sentence. The features include:

- For each event type, the average and minimum re-ranker confidence score, as well as the count of events of that type.
- For each event type, the count of events sharing the same trigger.
- For each event type, the count of events sharing the same arguments.
- Minimum and maximum confidence values of triggers and arguments in the TEES output for the sentence.
- The section in the article in which the sentence appears, as given in the ST data.

4.2 Training phase

To train the re-ranker and the regressor, false positive events are needed in addition to the true positive events in the training data. We thus apply TEES to the training data and train the re-ranker using the correct ranking of the extracted events. A true positive event is given the rank 1 and a false positive event gets the rank -1. A query structure is then defined, grouping all events from a single sentence to avoid mutual comparison of events across sentences and documents during the training phase.

The trained re-ranker is then again applied to the training data. For every sentence, the optimal threshold is set to be the re-ranker score of the last event which should be retained so as to maximize the F-score. In case the sentence only contains false positives, the highest score is used, increased by an empirically established value of 0.2. A similar strategy is applied for sentences only containing true positives by using the lowest score, decreased by 0.2.

In both steps, the SVM regularization parameter C is set by a grid search on the development set.

Applying TEES and the re-ranker back to the training set results in a notably smaller proportion of false positives than would be expected on a novel input. To obtain a fully realistic training dataset for the re-ranker and threshold regressor would involve re-training TEES in a cross-validation setting, but this was not feasible due to the tight schedule constraints of the shared task, and is thus left as future work.

	#	P	R	F
Simple events	833	-0.08	-0.36	-0.23
Protein mod.	191	+0.09	-2.09	-1.12
Binding	333	+0.43	-1.20	-0.44
Regulation	1944	+2.38	-0.67	+0.36
All	3301	+1.71	-0.73	+0.23

Table 6: Performance difference in percentage points against the TEES system in the official test set results, shown for different event types.

4.3 Error analysis

Although the re-ranking approach resulted in a consistent gain over the state-of-the-art TEES system on both the development and the test sets, the overall improvement is only modest. As summarized in Table 6, the gain over the TEES system can be largely attributed to regulation events which exhibit a 2.38pp gain in precision for a 0.67pp loss in recall. Regulation events are at the same time by far the largest class of events, thus affecting the overall score the most.

In this section, we analyse the re-ranker and threshold regressor in isolation to understand their individual contributions to the overall result and to identify interesting directions for future research.

To isolate the re-ranker from the threshold regressor and to identify the maximal attainable performance, we set an oracle threshold in every sentence so as to maximize the sentence F-score and inspect the performance at this threshold, effectively bypassing the threshold regressor. This, however, provides a very optimistic estimate for sentences where all predicted events are false positives, because the oracle then simply obtains the

All events	P	R	F
B-C oracle (re-ranked)	81.32	39.61	53.27
W-C oracle (re-ranked)	54.92	39.61	46.02
W-C oracle (random)	51.06	39.19	44.34
Current system	47.15	39.61	43.05
TEES	45.46	40.39	42.77
Single-arg. events			
B-C oracle (re-ranked)	81.37	50.58	62.38
W-C oracle (re-ranked)	56.09	50.58	53.19
W-C oracle (random)	52.73	50.00	51.33
Current system	48.66	50.44	49.53
TEES	47.16	51.09	49.04
Multiple-arg. events			
B-C oracle (re-ranked)	81.02	16.83	27.87
W-C oracle (re-ranked)	48.61	16.83	25.00
W-C oracle (random)	42.66	16.75	24.05
Current system	39.64	17.12	23.91
TEES	37.57	18.17	24.50

Table 7: Performance comparison of the best case (B-C) and worst case (W-C) oracles, the current system with the re-ranker and threshold regressor, and TEES. As an additional baseline, the worst case oracle is also calculated for randomly ranked output. All results are reported also separately for single and multiple-argument events.

decisions from the gold standard and the ranking itself is irrelevant. This effect is particularly pronounced in sentences where only a single, false positive event is predicted (15.9% of all sentences with at least one event). Therefore, in addition to this *best case* oracle score, we also define a *worst case* oracle score, where no events are removed from sentences containing only false-positives. This error analysis is carried out on the development set using our own implementation of the performance measure to obtain per-event correctness judgments.

The results are shown in Table 7. Even for the worst case oracle, the re-ranked output has the potential to provide a 9.5pp increase in precision for a 0.8pp loss in recall over the baseline TEES system. How much of this potential gain is realized depends on the accuracy of the threshold regressor. In the current system, only a 1.7pp precision increase for a 0.8pp recall loss is attained, demonstrating that the threshold regressor leaves much room for improvement.

The best case oracle precision is 26.4pp higher than the worst case oracle, indicating that substantial performance losses can be attributed to sentences with purely false positive events. Indeed, sentences only containing one or two incorrect events account for 26% of all sentences with at least one predicted event. Due to their large impact on the overall system performance, these cases may justify a focused effort in future research.

	TEES	1-arg	N-arg	Full
Simple events	64.43	+0.07	±0.00	+0.07
Protein mod.	40.47	+0.06	±0.00	+0.06
Binding	82.03	±0.00	±0.00	±0.00
Regulation	30.34	+0.70	-0.14	+0.53
All events	45.04	+0.66	±0.00	+0.64

Table 8: Performance of the system on the development set when applied to single-argument events only (*1-arg*), to multiple-argument events only (*N-arg*), and to all events (*Full*).

To establish the relative merit of the re-ranker, we compare the worst-case oracle scores of the re-ranked output against random ranking, averaged over 10 randomization runs. While the difference between TEES output and the random ranking reflects the effect of using an oracle to optimize per-sentence score, the difference between the random ranking and the re-ranker output shows an actual added value of the re-ranker, not attained from the use of oracle thresholds. Here it is of particular interest to note that this difference is more pronounced for events with multiple arguments (5.95pp of precision) as opposed to single-argument events (3.36pp of precision), possibly due to the fact that such events have a much richer feature representation and also employ the EVEX resource. To assess the contribution of EVEX data, a re-ranker was trained solely on features derived from EVEX. This re-ranker achieved an F-score of 1.26pp higher than randomized ranking, thus suggesting that these features have a positive influence on the overall score.

To verify these results and measure their impact on the official evaluation, Table 8 summarizes the performance on the development set using the official evaluation service. To study the effect on single-argument events (column *1-arg*), the re-ranker score for multiple-argument events is artificially increased to always fall above the threshold. A similar strategy is used to study the effect on multiple-argument events (column *N-arg*). These results confirm that the overall performance gain of our system on top of TEES is obtained on single-argument events. Further, multiple-argument events have only a negligible effect on the overall score, demonstrating that, due to their low frequency, little can be gained or lost purely on multiple-argument events.

To summarize the error analysis, the results in

Table 7 suggest that the re-ranker is more effective on multiple-argument events where it receives more features including external information from EVEX. On the other hand, the results in Table 8 clearly demonstrate that the system is overall more effective on single-argument events. This would suggest a "mismatch" between the re-ranker and the threshold regressor, each being more effective on a different class of events. One possible explanation is the fact that the threshold regressor predicts a single threshold for all events in a sentence, regardless of their type and number of arguments. If these cannot be distinguished by one threshold, it is clear that the threshold regressor will optimize for the largest event type, i.e. a single-theme regulation. Studying ways to allow the regressor to act separately on various event types will be important future work.

4.4 Discussion and future work

One of the main limitations of our approach is that it can only increase precision, but not recall, since it removes events from the TEES output, but is not able to introduce new events. As TEES utilizes separate processing stages for predicting event triggers and argument edges, recall can be adjusted by altering either of these steps. We have briefly experimented with modifying TEES to over-generate events by artificially lowering the prediction threshold for event triggers. However, this simple strategy of over-generating triggers leads to a number of clearly incorrect events and did not provide any performance gain. As future work, we thus hope to explore effective ways to over-generate events in a more controlled and effective fashion. In particular, a more detailed evaluation is needed to assess whether the rate of trigger over-generation should be adjusted separately for each event type. Another direction to explore is to over-generate argument edges. This will entail a detailed analysis of partially correct events with a missing argument in TEES output. As in the case of triggers, it is likely that each event type will need to be optimized separately.

A notable amount of sentences include only false positive predictions, severely complicating the threshold regression. In an attempt to overcome this issue, we trained a sentence classifier for excluding sentences that should not contain any events. This classifier partially utilized the same features as the threshold regressor, as well as bag of words and bag of POS tags. This method showed some promise when used together with trigger over-generation, but the gain was not enough to surpass the lost precision caused by the over-generation. If the event over-generation can be improved, the feasibility of this method should be re-evaluated.

5 Conclusions

We have presented our participation in the latest BioNLP Shared Task by mainly relying on the large-scale text mining resource EVEX. For the GRN task, we were able to produce a gene regulatory network from the EVEX data without retraining specific text mining algorithms. Using predicted gene/protein symbols and the GE formalism, rather than gold standard entities and the GRN annotation scheme, our final result on the test set only performed 0.06 SER points worse as compared to the corresponding TEES submission. This encouraging result warrants the use of generic large-scale text mining data in network biology settings. As future work, we will extend the EVEX dataset with information on the entity types to enable pruning of false-positive events and more fine-grained classification of event types, such as the distinction between promoter binding (Protein-Gene Binding) and protein-protein interactions (Protein-Protein Binding).

In the GE task, we explored a re-ranking approach to improve the precision of the TEES event extraction system, also incorporating features from the EVEX resource. This approach led to a modest increase in the overall F-score of TEES and resulted in the first rank on the GE task. In the subsequent error analysis, we have demonstrated that the re-ranker provides an opportunity for a substantial increase of performance, only partially realized by the regressor which sets a per-sentence threshold. The analysis has identified numerous future research directions.

Acknowledgments

Computational resources were provided by CSC IT Center for Science Ltd., Espoo, Finland. The work of KH and FG was supported by the Academy of Finland, and of SVL by the Research Foundation Flanders (FWO). YVdP and SVL acknowledge the support from Ghent University (Multidisciplinary Research Partnership Bioinformatics: from nucleotides to networks).

References

Jari Björne and Tapio Salakoski. 2013. TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task. In *Proceedings of BioNLP Shared Task 2013 Workshop*. In press.

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

Jari Björne, Filip Ginter, and Tapio Salakoski. 2012b. University of Turku in the BioNLP'11 Shared Task. *BMC Bioinformatics*, 13(Suppl 11):S4.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Suwisa Kaewphan, Sanna Kreula, Sofie Van Landeghem, Yves Van de Peer, Patrik Jones, and Filip Ginter. 2012. Integrating large-scale text mining and co-expression networks: Targeting NADP(H) metabolism in E. coli with event extraction. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012)*.

Jin-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.

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.

David McClosky, Sebastian Riedel, Mihai Surdeanu, Andrew McCallum, and Christopher Manning. 2012. Combining joint models for biomedical event extraction. *BMC Bioinformatics*, 13(Suppl 11):S9.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

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 Inzé, 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.

Andreas Vlachos and Mark Craven. 2012. Biomedical event extraction from abstracts and full papers using search-based structured prediction. *BMC Bioinformatics*, 13(Suppl 11):S5.

Chih-Hsuan Wei and Hung-Yu Kao. 2011. Cross-species gene normalization by species inference. *BMC Bioinformatics*, 12(Suppl 8):S5.

Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, and Yuji Matsumoto. 2011. Coreference based event-argument relation extraction on biomedical text. *Journal of Biomedical Semantics*, 2(Suppl 5):S6.