How Feasible and Robust is the Automatic Extraction of Gene Regulation Events ? A Cross-Method Evaluation under Lab and Real-Life Conditions

Udo Hahn[1] **Katrin Tomanek**[1] **Ekaterina Buyko**[1] **Jung-jae Kim**[2] **Dietrich Rebholz-Schuhmann**[2]

[1]Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Germany
{udo.hahn|katrin.tomanek|ekaterina.buyko}@uni-jena.de

[2]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
{kim|rebholz}@ebi.ac.uk

Abstract

We explore a rule system and a machine learning (ML) approach to automatically harvest information on gene regulation events (GREs) from biological documents in two different evaluation scenarios – one uses self-supplied corpora in a clean lab setting, while the other incorporates a standard reference database of curated GREs from REGULONDB, real-life data generated independently from our work. In the lab condition, we test how feasible the automatic extraction of GREs really is and achieve F-scores, under different, not directly comparable test conditions though, for the rule and the ML systems which amount to 34% and 44%, respectively. In the REGULONDB condition, we investigate how robust both methodologies are by comparing them with this routinely used database. Here, the best F-scores for the rule and the ML systems amount to 34% and 19%, respectively.

1 Introduction

The extraction of binary relations from biomedical text has caught much attention in the recent years. Progress on this and other tasks has been monitored in challenge competitions such as BIOCREATIVE I and II,[1] which dealt with gene/protein names and and protein-protein interaction.

The BIOCREATIVE challenge and other related ones have shown at several occasions that participants continue to use two fundamentally different systems: symbolic pattern-based systems (rule systems), on the one hand, and feature-based statistical machine learning (ML) systems, on the other hand. This has led to some rivalry with regard to the interpretation of their performance data, the costs of human efforts still required and their scalability for the various tasks. While rule systems are often hand-crafted and fine-tuned to a particular application (making a major manual rewrite often necessary when the application area is shifted), ML systems are trained automatically on manually annotated corpora, i.e., without manual intervention, and thus have the advantage to more easily adapt to changes in the requested identification tasks. Time costs (human workload) are thus shifted from rule design and adaptation to metadata annotation.

Text mining systems as usually delivered by BioNLP researchers render biologically relevant entities and relations on a limited set of test documents only. While this might be sufficient for the BioNLP community, it is certainly insufficient for bioinformaticians and molecular biologists since they require large-scale data with high coverage and reliability. For our analysis, we have chosen the topic of gene regulatory events in E. coli, which is a domain of very active research and grand challenges.[2] Currently the gold standard of the existing body of knowledge of such events is represented by the fact database REGULONDB.[3] Its content has been man-

[1]http://biocreative.sourceforge.net/

[2]The field of gene regulation is one of the most prominent topics of research and often mentioned as one of the core fields of future research in molecular biology (cf, e.g., the Grand Challenge I-2 described by Collins et al. (2003)).

[3]http://regulondb.ccg.unam.mx/

Proceedings of the Workshop on BioNLP, pages 37–45,
Boulder, Colorado, June 2009.

ually gathered from the scientific literature and describes the curated computational model of mechanisms of transcriptional regulation in E. coli. Having this gold standard in mind, we face the challenging task to automatically reproduce this content from the available literature, to enhance this content with reliable additional information and to update this resource as part of a regular automatic routine.

Hence, we first explore the *feasibility* and performance of a rule-based and an ML-based system against special, independently created corpora that were generated to enable measurements under clean experimental lab conditions. This part, due to different experimental settings, is not meant as a comparison between both approaches though. We then move to the even more demanding real-life scenario where we evaluate and compare these solutions for the identification of gene regulatory events against the REGULONDB data resource. This approach targets the *robustness* of the proposed text mining solutions from the perspectives of completeness, correctness and novelty of the generated results.

2 Related Work

Considering relation extraction (RE) in the biomedical domain, there are only few studies which deal primarily with gene regulation. Yang et al. (2008) focus on the detection of sentences that contain mentions of transcription factors (proteins regulating gene expression). They aim at the detection of new transcription factors, while relations are not taken into account. In contrast, Šarić et al. (2004) extract gene regulatory networks and achieve in the RE task an accuracy of up to 90%. They disregard, however, ambiguous instances, which may have led to the low recall around 20%. The *Genic Interaction Extraction Challenge* (Nédellec, 2005) was organized to determine the state-of-the-art performance of systems designed for the detection of gene regulation interactions. The best system achieved a performance of about 50% F-score. The results, however, have to be taken with care as the LLL corpus used in the challenge is of extremely limited size.

3 Extraction of Gene Regulation Events

Gene regulation is a complex cellular process that controls the expression of genes. These genes are then transcribed into their RNA representation and later translated into proteins, which fulfill various tasks such as maintaining the cell structure, enabling the generation of energy and interaction with the environment.

The analysis of the gene regulatory processes is ongoing research work in molecular biology and affects a large number of research domains. In particular the interpretation of gene expression profiles from microarray analyses could be enhanced using our understanding of gene regulation events (GREs) from the literature.

We approach the task of the automatic extraction of GREs from literature from two different methodological angles. On the one hand, we provide a set of hand-crafted rules – both for linguistic analysis and conceptual inference (cf. Section 3.1), the latter being particularly helpful in unveiling only implicitly stated biological knowledge. On the other hand, we supply a machine learning-based system for event extraction (cf. Section 3.2). No regularities are specified *a priori* by a human although, at least in the supervised scenario we have chosen, this approach relies on training data supplied by human (expert) annotators who provide sufficiently many instances of ground truth decisions from which regularities can automatically be learnt. At the level of system performance, rules tend to foster precision at the cost of recall and ML systems tend to produce inverse figures, while there is no conclusive evidence for or against any of these two approaches.

The extraction of GREs, independent of the approach one subscribes to, is a complex problem composed of a series of subtasks. Abstracting away from lots of clerical and infrastructure services (e.g., sentence splitting, tokenization) at the core of any GRE extraction lie the following basic steps:

- the identification of pairs of gene mentions as the arguments of a relation – the well-known named entity recognition and normalization task,
- the decision whether the entity pair really constitutes a relation,
- and the identification of the roles of the arguments in the relation which implicitly amounts to characterize each argument as either agent or patient.

3.1 Rule-based Extraction

The rule-based system extracts GREs from text employing logical inference. The motivation of using inference is that the events under scrutiny are often expressed in text in either a compositional or an incomplete way. We address this issue by compositionally representing textual semantics and by logically inferring implicit meanings of text over the compositional representation of textual semantics.

Entity Identification. The system first recognizes named entities of the types that can be participants of the target events. We have collected 15,881 E. coli gene/protein and operon names from REGULONDB and UNIPROT. Most of the gene/protein names are associated with UNIPROT identifiers. An operon in prokaryotes is a DNA sequence with multiple genes whose expression is controlled by a shared promoter and which thus express together. We have mapped the operon names to corresponding gene sets.

Named entity recognition relies on the use of dictionaries. If the system recognizes an operon name, it then associates the operon with its genes. The system further recognizes multi-gene object names (e.g., "acrAB"), divides them into individual gene names (e.g., "acrA", "acrB") and associates the gene names with the multi-gene object names.

Relation Identification. The system then identifies syntactic structures of sentences in an input corpus by utilizing the ENJU parser (Sagae et al., 2007). The ENJU parser generates predicate-argument structures, and the system converts them into dependency structures.

The system then analyzes the semantics of the sentences by matching syntactic-semantic patterns to the dependency structures. We constructed 1,123 patterns for the event extraction according to the following workflow. We first collected keywords related to gene regulation, from GENE ONTOLOGY, INTERPRO, WORDNET, and several papers about information extraction from biomedical literature (Hatzivassiloglou and Weng, 2002; Kim and Park, 2004; Huang et al., 2004). Then we collected subcategorization frames for each keyword and created patterns for the frames manually.

Each pattern consists of a syntactic pattern and a semantic pattern. The syntactic patterns comply with dependency structures. The system tries to match the syntactic patterns to the dependency structures of sentences in a bottom-up way, considering syntactic and semantic restrictions of syntactic patterns. Once a syntactic pattern is successfully matched to a sub-tree of the available dependency structure, its corresponding semantic pattern is assigned to the sub-tree as one of its semantics. The semantic patterns are combined according to the dependency structures to form a compositional semantic structure.

The system then performs logical inference over the semantic structures by using handcrafted inference rules and extracts target information from the results of the inference. We have manually created 28 inference rules that reflect the knowledge of the gene regulation domain. Only relations where the identified agent is one of those known TFs are kept, while all others are discarded.

3.2 Generic, ML-based Extraction

Apart from the already mentioned clerical preprocessing steps, the ML-based extraction of GREs requires several additional syntactic processing steps including POS-tagging, chunking, and full dependency- and constituency-based parsing.[4]

Entity Identification. To identify gene names in the documents, we applied GENO, a multi-organism gene name recognizer and normalizer (Wermter et al., 2009) which achieved a top-rank performance of 86.4% on the gene normalization task of BIOCREATIVE-II. GENO recognizes gene mentions by means of an ML-based named entity tagger trained on publicly available corpora. Subsequently, it attempts to map all identified mentions to organism-specific UNIPROT[5] identifiers. Mentions that cannot be mapped are discarded; only successfully mapped mentions are kept. We utilized GENO in its original version, i.e., without special adjustments to the E. coli organism. However, only those mentions detected to be genes of E. coli were fed into the relation extraction component.

[4] These tasks were performed with the OPENNLP tools (`http://opennlp.sourceforge.net/`) and the MST parser (`http://sourceforge.net/projects/mstparser`), both retrained on biomedical corpora.

[5] `http://www.uniprot.de`

Relation Identification. The ML-based approach to GRE employs Maximum Entropy models and constitutes and extension of the system proposed by Buyko et al. (2008) as it also makes use of dependency parse information including dependency tree level features (Katrenko and Adriaans, 2006) and shortest dependency path features (Kim et al., 2008). In short, the feature set consists of:

- word features (covering words before, after and between both entity mentions);
- entity features (accounting for combinations of entity types, flags indicating whether mentions have an overlap, and their mention level);
- chunking and constituency-based parsing features (concerned with head words of the phrases between two entity mentions; this class of features exploits constituency-based parsing as well and indicates, e.g., whether mentions are in the same NP, PP or VP);
- dependency parse features (analyzing both the dependency levels of the arguments as discussed by Katrenko and Adriaans (2006) and dependency path structure between the arguments as described by Kim et al. (2008));
- and relational trigger (key)words (accounting for the connection of trigger words and mentions in a full parse tree).

An advantage of ML-based systems is that they allow for thresholding. To achieve higher recall values for our system, we may set the confidence threshold for the negative class (i.e., a pair of entity mentions does not constitute a relation) to values > 0.5. Clearly, this is at the cost of precision as the system more readily assigns the positive class.

4 Intrinsic Evaluation of Feasibility

The following two sections aim at evaluating the rule-based and ML-based GRE extraction systems. The systems are first "intrinsically" evaluated, i.e., in a cross-validation manner on corpora annotated with respect to GREs. Second, in a more realistic scenario, both systems were evaluated against REGULONDB, a database collecting knowledge about gene regulation in E. coli. This scenario tests which part of manually accumulated knowledge about gene regulation in E. coli can automatically be identified by our systems and at what level of quality.

4.1 Rule-based system

Corpus. For the training and evaluation of the rule-based system, we annotated 209 MEDLINE abstracts with three types of events: specific events of gene transcription regulation, general events of gene expression regulation, and physical events of binding of transcription factors to gene regulatory regions. Strictly speaking, only the first type is relevant to REGULONDB. However, biologists often report gene transcription regulation events in the scientific literature as if they are gene expression regulation events, which is a generalization of gene transcription regulation, or the binding event, which itself is insufficient evidence for gene transcription regulation. The two latter types may indicate that the full-texts contain evidence of the first type.

We asked two curators to annotate the abstracts. Curator A was trained with example annotations and interactive discussions. Curator B was trained only with example annotations and guidelines. For cross-checking of annotations, we asked them to annotate an unseen corpus of 97 abstracts and found that Curator A made 10.8% errors, misjudging three event additions and, in the other 14 errors, mistaking in annotating event types, event attributes, and passage boundaries, while Curator B made 32.4% errors as such. This result indicates that the annotation of GREs requires intensive and interactive training. The curators have discussed and agreed on the final release of the corpora.[6]

Results. The system has successfully extracted 79 biologically meaningful events among them (21.1% recall) and incorrectly produced 15 events (84.0% precision) which constitutes an overall F-score of 33.6%. Among the 79 events, the system has correctly identified event types of 39 events (49.4% precision), polarity of 46 events (58.2% precision), and directness of 51 events (64.6% precision). Note that the system employed a fully automatic module for named entity recognition. The event type recognition is impaired, because it often fails to recognize

[6]The resultant annotated corpora are available at `http://www.ebi.ac.uk/~kim/eventannotation/`.

the specific event type of transcription regulation, but only identifies the general event type of gene expression regulation due to the lack of identified evidence.

4.2 ML-based system

GeneReg corpus. The GENEREG corpus (Buyko et al., 2008) constitutes a selection of 314 MEDLINE abstracts related to gene regulation in E. coli. These abstracts were randomly drawn from a set of 32,155 selected by MESH term queries from MEDLINE using keywords such as *Escherichia coli*, *Gene Expression* and *Transcription Factors*. These 314 abstracts were manually annotated for named entities involved in gene regulatory processes (such as transcription factor, including co-factors and regulators, and genes) and pairwise relations between transcription factors (TFs) and genes, as well as triggers (e.g., clue verbs) essential for the description of gene regulation relations. As for the relation types, the GENEREG corpus distinguishes between (a) unspecified regulation of gene expression, (b) positive, and (c) negative regulation of gene expression. Out of the 314 abstracts a set of 65 were randomly selected and annotated by a second annotator to identify inter-annotator agreement (IAA) values. For the task of correct identification of the pair of interacting named entities in gene regulation processes, an IAA of 78.4% (R), 77.3% (P), 77.8% (F) was measured , while 67% (R), 67.9% (P), 67.4% (F) were achieved for the identification of interacting pairs plus the 3-way classification of the interaction relation.

Experimental Setting. The ML-based extraction system merges all of the above mentioned three types (unspecific, negative and positive) into one common type "relation of gene expression". So, it either finds that there is a relation of interest between a pair of gold entity mentions or not. We evaluated our system by a 5-fold cross-validation on the GENEREG corpus. The fold splits were done on the abstract-level to avoid the otherwise unrealistic scenario where a system is trained on sentences from an abstract and evaluated on other sentences but from the same abstract (Pyysalo et al., 2008). As our focus here is only on the performance of the GRE extraction component, gold entity mentions as annotated in the respective corpus were used.

Results. For the experimental settings given above, the system achieved an F-score of 42% with a precision of 59% and a recall of 33%. Increasing the confidence threshold for the negative class increases recall as shown for two different thresholds in Table 1. As expected this is at the cost of precision. It shows, that using an extremely high threshold of 0.95 results in a dramatically increased recall of 73% compared to 33% with the default threshold. Although at the cost of diminished precision of 32% compared to originally 59%, the lifted threshold increases the overall F-score (44%) by 2 points.

threshold	R	P	F
default (0.5)	0.33	0.59	0.42
0.80	0.54	0.43	0.48
0.95	0.73	0.32	0.44

Table 1: Different confidence thresholds for the ML-based system achieved by intrinsic evaluation

5 Extrinsic Evaluation of Robustness

REGULONDB is the primary and largest reference database providing manually curated knowledge of the transcriptional regulatory network of E. coli K12. On K12, approximately for one-third of K12's genes, information about their regulation is available. REGULONDB is updated with content from recent research papers on this issue. While REGULONDB contains much more, for this paper our focus was solely on REGULONDB's information about gene regulation events in E. coli. In the following, the term REGULONDB refers to this part of the REGULONDB database. REGULONDB includes e.g., the following information for each regulation event: regulatory gene (the "agent" in such an event, a transcription factor), the regulated gene (the "patient"), the regulatory effect on the regulated gene (activating, suppression, dual, unknown), and evidence that supports the existence of the regulatory interaction.

Evaluation against REGULONDB constitutes a real-life scenario. Thus, the complete extraction systems were run, including gene name recognition and normalization as well as relation detection. Hence, the systems' overall recall values are highly affected by the gene name identification. REGULONDB is here taken as a "true" gold standard and thus as-

sumed to be correct and exhaustive with respect to the GREs contained. As, however, every manually curated database is likely to be incomplete and might contain some errors, we supplement our evaluation against REGULONDB with a manual analysis of false positives errors caused by our system (cf. Section 5.4).

5.1 Evaluation Scenario and Experimental Settings

To evaluate our extraction systems against REGULONDB we first processed a set of input documents (see below), collected all unique gene regulation events extracted and compared this set of events against the full set of known events in REGULONDB. A true positive (TP) hit is obtained, when an event found automatically corresponds to one in REGULONDB, i.e., having the same agent and patient. The type of regulation is not considered. A false positive (FP) hit is counted, if an event was found which does not occur in the same way in REGULONDB, i.e., either patient or agent (or both) are wrong. False negatives (FN) are those events covered by REGULONDB but not found by a system automatically. From these hit values, standard precision, recall, and F-score values are calculated. Of course, the systems' performance largely depend on the size of the base corpus collection processed. Thus, for both systems and all three document sets we got separate performance scores.

Table 2 gives an overview to the document collections used for evaluating the robustness of our systems: The "ecoli-tf" variants are documents filtered both with E. coli TF names and with relevance to E. coli. Abstracts are taken from Medline citations, while full texts are from a corpus of different biomedical journals. The third document set, "regulon ra", is a set containing abstracts from the REGULONDB references.

name	type	# documents
ecoli-tf.abstracts	abstract	4,347
ecoli-tf.fulltext	full texts	1,812
regulon ra	abstracts	2,704

Table 2: Document sets for REGULONDB evaluation

5.2 Rule-based-System

Table 3 shows the evaluation results of the rule-based system against REGULONDB. Though the system distinguishes the three types of events, we have considered them all as events of gene transcription regulation for the evaluation. For instance, the system has extracted 718 unique events with single-unit participants (i.e., excluding operons), not considering event types and attributes (e.g., polarity), from the "ecoli-tf.fulltext" corpus. Among the events, 347 events are found in Regulon (9.7% recall, 48.3% precision). If we only consider the events that are specifically identified as gene transcription regulation, the system has extracted 379 unique events among which 201 are also found in Regulon (5.6% recall, 53.0% precision).

participant	document set	R	P	F
single-unit	ecoli-tf.abstracts	0.09	0.60	0.15
multi-unit	ecoli-tf.abstracts	0.24	0.61	0.34
single-unit	ecoli-tf.fulltext	0.10	0.48	0.16
multi-unit	ecoli-tf.fulltext	0.25	0.49	0.33
single-unit	regulon ra	0.07	0.73	0.13
multi-unit	regulon ra	0.18	0.70	0.28

Table 3: Results of evaluation against REGULONDB of rule-based system.

When we split multi-unit participants into individual genes, the rule-based system shows better performance, as shown in Table 3 with the participant type "multi-unit". This may indicate that the gene regulatory events of E. coli are often described as interactions of operons. At best, the system shows 34% F-score with the "ecoli-tf.abstracts" corpus.

5.3 ML-based System

The ML-based system was designed to recognize all types of gene regulation events. REGULONDB, however, contains only the subtype, i.e., regulation of transcription. Thus, the ML-based system was evaluated against REGULONDB in two modes: by default, all events extracted by the systems are considered; in the "TF-filtered" mode, only relations with an agent from the list of all known TFs in E. coli are considered (as done for the rule-based system by default). Thus, comparing to the rule-based system, only the results obtained in the "TF-filtered" mode should be considered.

5.3.1 Raw performance scores

The results for the ML-based system are shown in Table 4. Recall values here range between 7 and 10%, while precision is between 29 and 78% depending on both the document set as well as the application of the TF filter. The low recall of the ML-based system is partially due to the fact that the system does not recognize multi-gene object names (e.g., "acrAB"), in this configuration the recall is similar to the recall of the rule-based system in a "single-unit modus" (see Table 3).

mode	document set	R	P	F
TF-filtered	ecoli-tf.abstracts	0.09	0.70	0.16
default	ecoli-tf.abstracts	0.09	0.45	0.15
TF-filtered	ecoli-relevant.fulltext	0.10	0.54	0.17
default	ecoli-relevant.fulltext	0.10	0.29	0.15
TF-filtered	regulon ra	0.07	0.78	0.13
default	regulon ra	0.07	0.47	0.12

Table 4: Results of evaluation against REGULONDB of ML-based system

As already shown in the intrinsic evaluation, application of different confidence thresholds increases the recall of the ML-based system. This was also done for the evaluation against REGULONDB. Table 5 shows the impact of increased confidence thresholds for the negative class on the "regulon ra" set for the "TF-filtered" evaluation mode. Given an extremely high threshold of 0.95, the recall is increased from 7 to 11% which constitutes a relative increase of over 60%. Precision obviously drops, however, the overall F-score has improved from 13 to 19%. These results emphasize that an ML-based system has an important handle which allows to adjust recall according to the application needs.

threshold	R	P	F
default (0.5)	0.07	0.78	0.13
0.8	0.09	0.70	0.16
0.95	0.11	0.63	0.19

Table 5: Different confidence thresholds for the ML-based system tested on the "regulon ra" set

5.4 Manual analysis of false positives

REGULONDB was taken as an absolute gold standard in this evaluation. If a system correctly extracts an event which is not contained in REGULONDB for some reason, this constitutes a FP. Moreover, all kinds of error (e.g., agent and patient mixed up) were subsumed as FP errors. To analyze the cause and distribution of FPs in more detail, a manual analysis of the FP errors was performed and original FP hits were assigned to one out of four FP error categories:

Cat1: Not a GRE This is really an FP error, as the extracted relation does not at all constitute a gene regulation event.

Cat2a: GRE but other than transcription Unlike REGULONDB which contains only one subtype of GREs, namely transcriptions, the ML-based system identifies all kinds of GREs. Therefore, the ML-based system clearly identifies events which cannot be contained in REGULONDB and, therefore, are not really FPs.

Cat 3: Partially correct transcription event This category deals with incorrect arguments of GREs. We distinguish three types of FPs: (a) the patient and the agent role are interchanged, (b) the patient is wrong, while the agent is right, and (c) the agent is wrong, while the patient is right. In all these three cases, though errors were committed human curators might find the partially incorrect information useful to speed up the curation process.

Cat4: Relation missing in REGULONDB Those are relations which should be contained in REGULONDB but are missing for some reason. The agent is a correctly identified transcription factor and the sentence contains a mention of a transcription event. There are several reasons why this relation was not found in REGULONDB as we will discuss in the following.

Table 6 shows the results of the manual FP analysis of the ML-based system (no TF filter applied) on the "ecoli-tf-abstracts" and "ecoli-tf-fulltexts". It shows that the largest source of error is due to `Cat1`, i.e., an identified relation is completely wrong. As fulltext documents are generally more complex, the relative amount of this kind of errors is higher here than on abstracts (54.5 % compared

category	abstracts (%)	fulltexts (%)
Cat 1	44.5	54.5
Cat 2	11.2	10.9
Cat 3a	3.8	3.9
Cat 3b	8.5	4.4
Cat 3c	8.2	5.4
Cat 4	23.8	21.0

Table 6: Manual analysis of false positive errors (FP). Percentages of FPs by category are reported on "ecoli-tf-abstracts" and "ecoli-tf-fulltexts"

to 44.5 %). However, on abstracts and fulltexts, a bit more than 10 % of the FP are because the system found too general GREs which, by definition, are not contained in REGULONDB (`Cat2`). Identified GREs that were partially correct constitute 20.5 % (abstracts) or 13.7 % (fulltexts) of the FP errors (`Cat3`).

Finally, 23.8% and 21.0% of the FPs for abstracts and fulltext, respectively, are correct transcription events but could not be found in REGULONDB (`Cat4`). This is due to several reasons. For instance, identified gene names were incorrectly normalized so that they could not be found in REGULONDB, REGULONDB curators have not yet added a relation or simply overlooked it; relations are correctly identified as such in the narrow context of a paragraph of a document but were actually of speculative nature only (this includes relations whose status is unsure, often indicated by "likely" or "possibly").

Summarizing, the manual FP analysis shows that about 50% of all FPs are not completely erroneous. These numbers must clearly be kept in mind when interpreting the raw numbers (especially for precision) reported on in the previous subsection.

5.5 Integration of text mining results

We have integrated the results of the two different text mining systems and found that both systems are complementary to each other such that their result sets do not heavily overlap. For instance, from the "ecoli-tf.abstract" corpus, the rule-based system extracts 992 events, while the ML-based system extracts 705 events. For the integration, we have considered only the events whose participants are associated with UNIPROT identifiers. Among the extracted events, only 285 events are extracted by both systems. We might speculate that the overlapping events are more reliable than the rest of the extracted events. It also leaves 71.3% of the results from the rule-based system and 59.6% of results from the ML-based system as unique contributions from each of the approaches for the integration.

6 Conclusions

We have explored a rule-based and a machine learning-based approach to the automatic extraction of gene regulation events. Both approaches were evaluated under well-defined lab conditions using self-supplied corpora, and under real-life conditions by comparing our results with REGULONDB, a well-curated reference data set. While the results for the first evaluation scenario are state of the art, performance figures in the real-life scenario are not so shiny (the best F-scores for the rule-based and the ML-based system are on the order of 34% and 19%, respectively). This holds, in particular, for the comparison with the work of Rodríguez-Penagos et al. (2007). Still, at least the ML-based approach is much more general than the very specifically tuned manual rule set from Rodríguez-Penagos et al. (2007) and has potential for increases in performance. Also, this has been the first extra-mural evaluation of automatically generating content for REGULONDB.

Still, the analysis of false positives reveals that the strict criteria we applied for our evaluation may appear in another light for human curators. Confounded agents and patients (21% on the abstracts, 14% on full texts) and information not contained in REGULONDB (24% on the abstracts, 21% on full texts) might be useful from a heuristic perspective to focus on interesting data during the curation process.

Acknowledgements

This work was funded by the EC within the BOOTStrep (FP6-028099) and the CALBC (FP7-231727) projects. We want to thank Tobias Wagner (Centre for Molecular Biomedicine, FSU Jena) for performing the manual FP analysis.

References

Ekaterina Buyko, Elena Beisswanger, and Udo Hahn. 2008. Testing different ACE-style feature sets for the extraction of gene regulation relations from MEDLINE abstracts. In *Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, pages 21–28.

Francis Collins, Eric Green, Alan Guttmacher, and Mark Guyer. 2003. A vision for the future of genomics research. *Nature*, 422(6934 (24 Feb)):835–847.

Vasileios Hatzivassiloglou and Wubin Weng. 2002. Learning anchor verbs for biological interaction patterns from published text articles. *International Journal of Medical Informatics*, 67:19–32.

Minlie Huang, Xiaoyan Zhu, Donald G. Payan, Kunbin Qu, and Ming Li. 2004. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

Sophia Katrenko and P. Adriaans. 2006. Learning relations from biomedical corpora using dependency trees. In *KDECB 2006 – Knowledge Discovery and Emergent Complexity in Bioinformatics*, pages 61–80.

Jung-jae Kim and Jong C. Park. 2004. BioIE: retargetable information extraction and ontological annotation of biological interactions from the literature. *Journal of Bioinformatics and Computational Biology*, 2(3):551–568.

Seon-Ho Kim, Juntae Yoon, and Jihoon Yang. 2008. Kernel approaches for genic interaction extraction. *Bioinformatics*, 24(1):118–126.

Claire Nédellec. 2005. Learning language in logic - genic interaction extraction challenge. In *Learning language in logic - genic interaction extraction LLL' 2005*, pages 31–37.

Sampo Pyysalo, Antti Airola, Juho Heimonen, Jari Björne, Filip Ginter, and Tapio Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(3), April.

Carlos Rodríguez-Penagos, Heladia Salgado, Irma Martínez-Flores, and Julio Collado-Vides. 2007. Automatic reconstruction of a bacterial regulatory network using natural language processing. *BMC Bioinformatics*, 8(293).

Kenji Sagae, Yusuke Miyao, and Junichi Tsujii. 2007. HPSG parsing with shallow dependency constraints. In *Annual Meeting of Association for Computational Linguistics*, pages 624–631.

Jasmin Šarić, Lars J. Jensen, Rossitza Ouzounova, Isabel Rojas, and Peer Bork. 2004. Extracting regulatory gene expression networks from pubmed. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 191, Morristown, NJ, USA. Association for Computational Linguistics.

Joachim Wermter, Katrin Tomanek, and Udo Hahn. 2009. High-performance gene name normalization with GeNo. *Bioinformatics*, 25(6):815–821.

Hui Yang, Goran Nenadic, and John Keane. 2008. Identification of transcription factor contexts in literature using machine learning approaches. *BMC Bioinformatics*, 9(Supplement 3: S11).