Testing Different ACE-Style Feature Sets for the Extraction of Gene Regulation Relations from MEDLINE Abstracts

Ekaterina Buyko **Elena Beisswanger** **Udo Hahn**

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena
D-07743 Jena, Germany
{ekaterina.buyko|elena.beisswanger|udo.hahn}@uni-jena.de

Abstract

Based on an introspective manual analysis how relations between biomedical entities surface literally in scientific abstracts, we investigate the performance of different feature sets for biomedical relation extraction in a supervised machine-learning setting. We start from fairly simple ACE-style ones and increasingly include domain-specific knowledge in these feature sets. This turns out to have beneficial effects on the extraction performance of the system under scrutiny.

1 Introduction

While systems for the recognition and interpretation of named entities have reached, by and large, a stable performance plateau at the 80% level, the extraction of relations between these entities lags far behind these figures. In the newswire domain, e.g., the *Automatic Content Extraction Program* (ACE) (Doddington et al., 2004) features the best system with 36.8% f-score in the detection of relation mentions.[1] This data is even underperfomed by the winning system of the *BioCreative 2 Protein Interaction Sub-Task* (IPS) (Hirschman et al., 2007), whose performance results settled at only 28.8% f-score. Although for both competitions strict real-world requirements were imposed on the task – the recognition and interpretation of all named entities involved, plus the recognition and interpretation of the associated relation (and, for the biomedical domain, the mapping of entities onto unique database identifiers) – relation extraction remains a challenging research problem under any conceivable conditions.

Our approach to deal with the challenges arising from relation extraction (RE) in the biomedical domain is, first, to explore the possible reasons for the inherent hardness of this task through introspective manual text analysis. In Section 3 we discuss empirical phenomena underlying relation encodings in biomedical documents, including a large variety of patterns and reliance on inferential processes. We then consider the performance of a feature-based learning approach for RE. Since supervised machine learning relies on carefully crafted feature sets, we consider in Section 5 different varieties of these sets, starting from ones which have already proven useful in the newswire domain. We then explore, in a preliminary way though, the possible contribution of domain-specific features for further fine-tuning these feature sets. Encouraging as these results may be our current corpus (cf. Section 4) suffers from an unbalanced occurrence of (too few) positive examples (clearly an issue that has to be addressed in future work).

2 Related Work

As far as the state of the art in biomedical RE is concerned, the simplest method for the extraction of relations between named entities is the detection of bag-of-word-style *co-occurrences* of entities of interest within documents or sentences (e.g., Jenssen et al. (2001)). Co-occurrence-based approaches are characterized by a high recall at the cost of an extremely low precision. Furthermore, the type and direction of relation usually cannot be determined. RE approaches that focus on higher precision but often suffer from weaker recall are based on *manually defined patterns* (e.g., Blaschke et al. (1999)). Some pattern-based

[1] Results are published at http://www.nist.gov/speech/tests/ace/2007/

approaches exploit morpho-syntactic and syntactic information and are based on *automatically learning RE patterns* from large corpora (e.g, Hakenberg et al. (2005), Huang et al. (2004)). These methods provide higher recall than those based on manually defined patterns. *Rule-based RE approaches* typically exploit full parse data of sentences and additional semantic information (e.g., Yakushiji et al. (2001), Saric et al. (2004), Fundel et al. (2007)).

In the newswire domain, supervised approaches currently dominate RE. This is partly due to the availability of large annotated corpora such as ACE (Doddington et al., 2004) that can be used for the training of machine learning models, using, e.g., Support Vector Machines (SVM) or Maximum Entropy (ME) models. The systems either exploit SVM kernels especially designed for the comparison of syntactic trees (e.g., Zelenko et al. (2003), Bunescu and Mooney (2005)) or they incorporate a variety of lexical, morpho-syntactic and syntactic features (e.g., Kambhatla (2004), Zhou et al. (2005)).

Considering RE in the biomedical domain, to the best of our knowledge, there are few studies which deal exclusively 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, Saric et al. (2004) extract gene regulatory networks and achieve in the RE task an accuracy of up to 90%. They disregard, however, ambiguous instances that potentially lower recall (no recall measures are reported in this work). 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 LLL corpus which was especially created for the *Genic Interaction Extraction Challenge*, the AIMED[2] and the BIOINFER corpus (Pyysalo et al., 2007) are considered as the standard corpora in the biomedical domain for evaluation of RE performance (Pyysalo et al., 2008). They are usually used for the development and evaluation of feature-based systems (e.g., Katrenko and Adriaans (2006), Sætre et al. (2007)). Katrenko and Adriaans (2006) report on experiments in which they reached f-score peaks on the AIMED corpus with 54.3% and on the LLL corpus with 72.4%, respectively. It should be noted that the experimental settings of the evaluations are not always clear. Therefore, Airola et al. (2008) suggested to indicate the evaluation settings of the instance-, sentence- or document-wise evaluation and proposed to use the latter as a default evaluation setting.

In our work we decided to use supervised machine learning models for gene regulation data and investigate whether the experience gained in the newswire domain can be applied to the biomedical domain as well. We chose the system presented by Zhou et al. (2005) that achieved competitive results of 74.7% in RE on ACE.

3 Textual Patterns for Gene Regulation

Very briefly, the regulation of gene expression can be described as the process that modulates the frequency, rate or extent of gene expression, where gene expression is the process in which the coding sequence of a gene is converted into a mature gene product or products, namely proteins or RNA (taken from the definition of the Gene Ontology class *Regulation of Gene Expression*, GO:0010468).[3] Transcription factors, cofactors and regulators are proteins that play a central role in the regulation of gene expression.

To get acquainted with the textual appearance of gene regulation relations, we manually analyzed 50 randomly extracted sentences from MEDLINE abstracts (see Section 4 for a detailed description of that collection) in order to find recurrent patterns by which gene regulation relations are literally expressed in sublanguage documents. We found 58 encoded relations and discovered in this set at least thirteen stable patterns how gene regulation relations surface in texts. In the following list we rank these patterns by their frequency in the set (in descending order).

1. [`Agent`] V-active [`Patient` Action-NN]
   "IclR also represses the expression of iclR"
2. [`Patient` Action-NN] V-passive [`Agent`]
   "yeiL expression is positively activated by Lrp"

[2] `ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/`

[3] `http://www.geneontology.org/`

3. [Agent] - Action-JJ [Patient]
   “SlyA-induced proteins”
4. [Agent] *is an* Actor-NN *for* [Patient]
   “IclR is a repressor for the Escherichia coli aceBAK operon”
5. [Agent Action-NN] V-active [Patient Action-NN]
   “Elevation of ppGpp levels in growing cells ... triggered the induction of all usp genes.”
6. [Agent] V-active (*bind to*) [Patient] *promoter / site*
   “ZntR is a trans-acting repressor protein that binds to the znt promoter region”
7. [Agent] *is required / essential / involved in* [Patient Action-NN]
   “rpoS function is essential for bgl silencing”
8. [Action-NN *of* Patient] *by* [Agent]
   “transcription repression of the Escherichia coli acetate operon by IclR”
9. [Patient Action-NN] V-active [Agent]
   “Expression of the tau and ssu genes requires the LysR-type transcriptional regulatory proteins CysB and Cbl”
10. *Promoter of* [Patient] *contains binding site for* [Agent]
    “The promoters of the mar/sox/rob regulon of Escherichia coli contain a binding site (marbox) for the homologous transcriptional activators MarA, SoxS and Rob.”
11. [Patient Action-NN] V-passive (*caused by*) [Agent]
    “bgl silencing caused by C-terminally truncated H-NS”
12. [Agent Action-NN] V-active (*cause*) [Patient Action-NN]
    “Disruption of cueR caused loss of copA expression”
13. [Action-NN Patient] *is under control of* [Agent]
    “Synthesis of Cbl itself is under control of the CysB protein”

The two most frequent patterns that contain the mention of regulation verbs (V-active, V-passive) cover a large amount of relation instances (1. (15 relations), 2. (11 relations)). The use of adjectives and Actor nominalizations (Actor-NN such as *‘regulator’*) are other frequent patterns in the expression of gene regulation relations (3. (6 relations), 4. (4 relations)). Uncertain expressions such as *‘be essential’*, *‘be involved in’* are used for the description of unspecified gene regulation relations (7. (3 relations)). Besides regulation verbs and their nominalizations and adjectives indicating requirements or dependencies, authors frequently describe the molecular process of the binding of a transcrition factor to a gene region (promoter) or provide information that a gene contains a binding site for a transcription factor (6. (3 relations), 10. (1 relation)). Another pattern group contains causal relations between molecular processes in which gene and transcription factors are involved (5. (3 relations), 11. (1 relation), 12. (1 relation)). Five out of 58 relations could not be classified into the featured thirteen patterns.

We see two challenges for the correct detection of gene regulation relations in sentences (given adequate syntactic structures are available): (1) the detection of *is-a* relations between mentions of entities participating in regulation process, and (2) inferential processes on biomedical knowledge for correctly distinguishing between positive and negative regulation processes.

The first issue concerns the frequent occurrence of appositions and predicate noun relations between participants, and the use of anaphoric mentions of participants in the regulation process. For example, in *“zntR gene encodes a putative regulatory protein that controls the expression of the znt operon”* we have an *is-a* relation between *‘zntR gene’* and *‘putative regulatory protein’*.

The second challenging issue concerns the correct detection of the category of gene regulation relations (*positive*, *negative* or *unspecified*). Experimental environments for gene regulation detection often involve genetic modifications of transcription factors. By means of these genetic modifications and the expression levels of other genes, researchers implicitly draw conclusions about the role of the transcription factor in the gene regulation processes. The sentence *“The production of C51 microcin decreased or was absent in rpoS, crp and cya mutant cells.*” contains a description of the decrease of *‘C51’* expression level. The fact that *‘rpoS’*, *‘crp’* and *‘cya’* genes are inactivated (*‘mutant cells’*) leads to the conclusion that they positively regulate *‘C51’*. The detection of the correct type of relation requires knowledge about experimental conditions (indicated here by *‘mutant cells’*) and inferences on biomedical knowledge.

4 Corpus Annotation

In this section we introduce the *JULIE Lab GeneReg corpus* which consists of MEDLINE abstracts dealing with gene regulation in *E. Coli*. It provides three types of semantic annotations:

- named entities involved in gene regulatory processes, such as TFs (transcription factors, cofactors and regulators) and genes,
- pairwise relations between TFs and genes,
- triggers (e.g., clue verbs) essential for the description of gene regulation relations

For all three annotation levels the annotation vocabulary was taken from the *Gene Regulation Ontology* (GRO) (Beisswanger et al., 2008). GRO describes gene regulation processes occurring on the intra-cellular level (such as the binding of transcription factors to DNA binding sites) and physical entities that are involved in these processes (such as genes and transcription factors).

A set of 32,155 abstracts was compiled from MEDLINE based on a query including the MESH terms *Escherichia coli*, *Gene Expression* and *Transcription Factors* (amongst others). From this set we randomly selected a corpus of 314 abstracts for manual annotation.

4.1 Named Entities

All abstracts in the corpus were annotated manually by a graduate student of biology considering the semantic categories enumerated in Table 1 which also gives the annotation counts (for definitions of the selected categories see GRO).[4]

Named Entity Category	Annotations
Transcription Factor	2496
Transcription Cofactor	14
Transcription Regulator	40
Gene	2547
Gene Group	1180
Gene (anaphoric)	24
Gene Group (anaphoric)	71

Table 1: Number of entity annotations per semantic category

To assess the Inter-Annotator Agreement (IAA) for the entity annotation a second graduate student of biology annotated a subset of 248 abstracts. For this subset the IAA was computed applying three standard IAA measures for the NE task: Strict IAA (69% (R), 62% (P), 65% (F)), Correct-Span IAA (74% (R), 76% (P), 72% (F)) and Correct-Category IAA (79% (R), 81% (P), 80% (F)).

Additionally, the biologist annotated anaphoric mentions of *Gene* and *Gene Group* entities (see also Table 1). The motivation of this annotation task was to provide more instances of entity mentions for the annotation of gene regulation relations. Anaphoric mentions were only annotated in sentences containing gene regulation relations.

4.2 Relations

The corpus of MEDLINE abstracts as described in Section 4 was also annotated with relations by a graduate biologist in a two-step annotation process. In a first step, trigger words indicating mentions of gene regulation processes were annotated. In a second step, the relations between genes and TFs (affecting the expression of the gene) were annotated. Next, we describe the two-step annotation process in more detail.

4.2.1 Annotation of Trigger Words

In preparation of the trigger word annotation, one biologist and one linguist manually screened the abstracts in the corpus and compiled a list of verbalizations of molecular processes that frequently occurred in the description of gene regulation relations. These processes were grouped in five categories enumerated in Table 2 based on conceptualizations and definitions from the GRO.

Trigger words indicating textual mentions of the listed processes were annotated with the corresponding categories. A trigger is any literal verbal form that clearly signals the occurrence of a particular molecular process. Trigger words are basically main verbs, verb nominalizations and adjectives. For example, the sentence *"H-NS and*

Semantic Category	Annotations
GeneExpression	495 (15)
TranscriptionOfGene	46 (12)
RegulationOfGeneExpression (unspecified)	896 (82)
PositiveRegulationOfGeneExpression	835 (110)
NegativeRegulationOfGeneExpression	441 (93)

Table 2: Number of trigger word annotations per semantic category (unique annotations are in brackets)

[4] `http://www.ebi.ac.uk/Rebholz-srv/GRO/GRO.html`

StpA proteins stimulate expression of the maltose regulon in Escherichia coli." contains two trigger words: first, *'stimulate'* is a trigger for a process in the category *PositiveRegulationOfGeneExpression*, second, *'expression'* is a trigger for a process belonging to the category *GeneExpression*.

4.3 Annotation of Gene Regulation Relations

In the second step, the graduate biologist annotated pairwise relations between genes and transcription factors, cofactors and general regulators affecting the expression of the gene. This annotation was based on the Gro class *RegulationOfGeneExpression* with its two sub-classes *PositiveRegulationOfGeneExpression* and *NegativeRegulationOfGeneExpression*. The concept *RegulationOfGeneExpression (unspecified)* was used for the annotation of gene regulation processes that could not be specified as either a positive or a negative regulation. We chose single sentences as annotation context for this task so that only those textual mentions of gene regulation relations and their participants were annotated which occurred within the same sentence.

A relation instance contains two arguments, *Arg1* and *Arg2*. *Arg1* is occupied by the *agent*, i.e., the entity that plays the role of modifying the gene expression. *Arg2* is occupied by the *patient*, i.e., the entity of which the expression is modified. While agents are proteins that regulate the expression of genes, patients are typically the genes of which the expression is regulated by the agent. The sentence *"The uxuAB operon is negatively controlled by the uxuR and exuR regulatory gene products.*" denotes a *NegativeRegulationOfGeneExpression* relation between the gene *'uxuAB'* and two transcription factors, viz. *'uxuR'* and *'exuR'*.

A set of 65 randomly selected abstracts was annotated by the second graduate student of biology for determing the IAA. A Strict IAA of 82% (R), 84% (P), 83% (F) was achieved for the task of the trigger annotation. An IAA of 78.4% (R), 77.3% (P), 77.8% (F) was measured for the task of correct identification of the pair of interacting named entities in gene regulation processes, 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.

Semantic Category	Annotations
RegulationOfGeneExpression (unspecified)	408
PositiveRegulationOfGeneExpression	455
NegativeRegulationOfGeneExpression	272

Table 3: Number of gene regulation relation annotations per semantic category

5 Methods

The patterns discussed in Section 3 already reveal the diversity how gene regulation relations surface literally in texts. For the automatic extraction of gene regulation relations, we pursue a feature-based approach to RE that incorporates diverse lexical, syntactic and semantic features. We, first, selected features that had already proved useful in the detection of relationships between entities in the newswire domain and were evaluated on the Ace RE corpus (Doddington et al., 2004). These features were intensively explored in the work of Zhou et al. (2005). As a classification model we chose the Maximum Entropy model implementation in Mallet.[5]

5.1 Features for Relation Extraction

Zhou et al. (2005) investigated eight classes of features suited for RE: words, entity type, mention level, overlap, base phrase chunking, dependency tree, parse tree and semantic resources. We chose seven classes of features (excluding semantic resources that were suited only for the newswire domain). In the following, we will briefly introduce these features. (for more detailed information, see Zhou et al. (2005)). As a substitute for Zhou et al.'s semantic resources, we incorporated a semantic feature class that exploits information about trigger word occurrence in the sentence (in the full parse tree path).

In the following we distinguish between two entity mentions in pairwise relations, i.e., E1 and E2. E1 is the entity mention that occurs first in the sentence (before E2). If one of the mentions includes another entity mention, then the entity mention with a larger span is classified as E1.

5.1.1 Words Features

This feature class covers four categories of words: (1) the words of both entity mentions, (2)

[5] Mallet is available at `http://mallet.cs.umass.edu/index.php/Main_Page`

the words between the entity mentions, (3) the words before E1, (4) the words after E2. For both mentions, head words and their combination are considered. The window for the words before E1 and after E2 has a size of two words.

5.1.2 Entity Features

Entity features account for combinations of entity types, flags indicating whether mentions have an overlap, and their mention level. For the latter, we distinguish between name and anaphoric mentions.

5.1.3 Base Phrase Chunking Features

The chunking features are concerned with the head words of the phrases between the two entity mentions. Zhou et al. (2005) show that shallow parsing features play a critical role for RE. This feature class covers four categories of phrases: (1) the phrases of both entity mentions, (2) the phrases between the entity mentions, (3) the phrases before E1, (4) the phrases after E2.

5.1.4 Full Parsing Features

This class of features deals with full parse tree information. We included in our work five from eight features presented by Zhou et al. (2005) excluding dependency features that concern dependencies derived from full parse tree as they were unreproducible from the paper's descriptions. We selected, however, features that exploit constituency-based parsing and indicate whether mentions are in the same noun, prepositional or verbal phrase. The path of phrase labels (without duplicates) between two entity mentions and the path enriched with phrase head information are considered as well.

5.1.5 Relational Trigger Words and Keywords

This newly added feature class accounts for the connection of trigger words and mentions in a full parse tree. We exploit features indicating whether the top phrase in the parse path between the entity mentions contains a *RegulationOfGeneExpression* trigger or one of its sub-type triggers, and whether *TranscriptionOfGene/GeneExpression* triggers occur in the same noun phrase as entity mentions. To acount for the influence of the experimental intervention context on the proper detection of the gene regulation relation, we checked whether keywords describing experimental interventions (e.g., *'mutant'*, *'deletion'*) (altogether, 56 keywords) co-occur in the same noun phrase with E1 or E2. For the evaluation of these features on the AIMED corpus we compiled a dictionary of interaction event triggers from our regulation trigger list and terms used by Fundel et al. (2007).

6 Experiments and Results

For the evaluation of our feature-based approach to gene regulation RE, we performed a ten-fold sentence-wise cross-validation on the GENEREG and the AIMED corpus. For the evaluation of the RE task we used original annotations of named entities and relational trigger words (the AIMED corpus was automatically tagged for event trigger words (cf. Section 5.1.5)). Anaphoric mentions of entities were included in the evaluation as well. We considered thus only the detection of a gene regulation relation between two entity mentions. As the regulation relation is an asymmetric one, we distinguish between *ARG1-relation-ARG2* and *ARG2-relation-ARG1*, i.e., the order in which the entities appear in the sentence. The overall evaluation results reflect the mean of both relations. The ten-fold cross-validation was done for (1) the binary classification of *RegulationOfGeneExpression* relation and (2) the 3-way classification *PositiveRegulationOfGeneExpression*, *NegativeRegulationOfGeneExpression*, and *RegulationOfGeneExpression (unspecified)*.[6]

The GENEREG corpus was enriched with morpho-syntactic and syntactic information. For POS tagging, chunking and parsing we used the re-trained OPENNLP tool suite.[7] These tools had previously been re-trained (Buyko et al., 2006) on the GENIA corpus (Ohta et al., 2002).

We evaluated on two feature sets: *Feature Set 1* (Zhou et al., 2005) and *Feature Set 2* that contains, in addition, the features exploiting relational triggers (see 5.1.5). The evaluation results clearly indicate that the straightforward porting of RE feature types from the newswire domain to the specialized biomedical domain does not provide fully satisfactory results (see Table 4). The addition of the domain-specific features (relational trigger words) increases the performance by 3 percentage points for the detection of the generic gene regulation relation (63.0%), and by nearly 14 percentage points for the detection of the specific

[6] *RegulationOfGeneExpression* does not contain specifications as to whether it is positive or negative.

[7] `http://opennlp.sourceforge.net/`

Semantic Category	Feature Set 1			Feature Set 2		
	R	P	F	R	P	F
RegulationOfGeneExp. (generic)	50.0	76.2	60.0	55.0 (78.6)	75.4 (56.8)	63.0 **(65.5)**
RegulationOfGeneExp. (pos.)	28.9	60.6	37.2	39.1 (56.2)	66.0 (48.8)	47.3 (**50.0**)
RegulationOfGeneExp. (neg.)	18.0	51.8	24.4	29.5 (41.2)	60.5 (51.3)	37.6 (**44.5**)
RegulationOfGeneExp. (unspec.)	10.8	31.7	14.9	28.5 (45.6)	60.9 (47.3)	37.3 (**44.6**)
Overall (pos./neg./unspec. at once)	19.5	55.6	28.2	31.5 (46.6)	65.0 (49.3)	42.0 **(47.3)**
AIMed	42.5	66.8	**51.5**	42.9	64.8	51.3

Table 4: Results of Gene Regulation Relation Extraction on the GeneReg (lines 3–7) and AIMed (line 8) corpus (reduction of negative examples (under-sampling) in brackets)

gene regulation relations (42.0%). Surprisingly, the evaluation on the AIMed corpus reveals that the incorporation of such semantic features does not enhance the over-all performance.

The error analysis with respect to specific relations revealed that the relation representations covered by the most frequent patterns (see Section 3) are correctly detected. Still, the main trouble here is the incorrect analysis of coordinated structures by the OpenNLP parser. In the incorrectly parsed sentence *"NarL Expression from the Escherichia coli nrf operon promoter is activated by the anaerobically triggered transcription factor, FNR, and by the nitrate/nitrite ion-controlled response regulators, NarL or NarP, but is repressed by the IHF and Fis proteins."* only one out of the five relation mentions was detected (the relation between *'nrf'* and *'FNR'*).

Another prominent failure source is in the frequent occurrence of anaphoric expressions within sentences in abstracts. The sentence *"The expression of the appY gene is induced immediately by anaerobiosis, and this anaerobic induction is independent of Fnr, and AppY, but dependent on ArcA"* contains an anaphoric expression *"this anaerobic induction"* that is crucial for the detection of the relation between transcription factors and the induced gene.

Furthermore, the rather weak results for the classification of specific gene regulation relations are partly due to inferences required for the proper detection of the category of the gene regulation relations. Currently, we handle these inferences (inappropriately though) only by exploiting keywords in the features. In order to more adequately deal with this problem we will focus, in future work, on a hybrid approach which includes at least a modest level of inferential capabilities.

One of the problems we see in the training data is the severe imbalance of positive and negative instances. The corpus contains about 9,000 negative instances and 1,135 positive instances only. Such an imbalance may cause serious learning problems, and is reflected already by the low performance (e.g, the particularly low recall of the minority class). If the classifier is uncertain, it typically predicts the majority class (in our case, the negative class label).

The sparseness of positive learning examples can be reduced by balancing the number of positive and negative examples in the training data. Two sampling schemes are usually applied here, over-sampling and under-sampling. The aim of over-sampling is to increase the number of the minority class instances, the goal of under-sampling is to reduce the number of the majority class instances. In our experiments, we chose under-sampling by randomly reducing negative examples in the training data down to the double number of the positive instances. The evaluation results for under-sampling reveal a substantial gain in the f-score up to 5 percentage points (see Table 4). In these runs we achieved the best performance of 65.5% for the detection of the generic gene regulation relation, and 47.3% for the extraction of specific gene regulation relations.

7 Conclusions and Future Work

We presented here a supervised approach to relation extraction in the biomedical sub-domain of gene regulation. The main contributions of this paper are in the descriptive analysis of the inherent hardness of this task and in tests of different feature sets for the extraction of gene regulation relations. Our evaluation results reveal that the straightforward porting of feature types that have already proven useful in the newswire RE should be fine-tuned by integrating domain-specific feature sets. The balancing of the training data and a comparison of our approach with state-of-the art rule-based systems (e.g., Fundel et al. (2007)) are the focus of our future work.

Acknowledgments

This work was partly funded by the EC within the BOOTSTREP project (FP6-028099) under the 6th Framework Program.

References

A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of the BioNLP Workshop at ACL 2008*, pages 1–9. Columbus, Ohio, USA, June 19, 2008.

E. Beisswanger, V. Lee, J. Kim, D. Rebholz-Schuhmann, A. Splendiani, O. Dameron, S. Schulz, and U. Hahn. 2008. Gene Regulation Ontology (GRO): Design principles and use cases. In *Proceedings of the MIE 2008 Conference*, pages 9–14. Gøteborg, Sweden, 25-28 May 2008.

C. Blaschke, M. Andrade, C. Ouzounis, and A. Valencia. 1999. Automatic extraction of biological information from scientific text: Protein-protein interactions. In *Proceedings of the ISMB'99*, pages 60–67. Heidelberg, Germany, August 6-10, 1999.

R. Bunescu and R. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the HLT-EMNLP'05*, pages 724–731. Vancouver, B.C., Canada, October 6-8, 2005.

E. Buyko, J. Wermter, M. Poprat, and U. Hahn. 2006. Automatically adapting an NLP core engine to the biology domain. *Proceedings of the Joint Bio-Ontologies and BioLINK Meeting at ISMB 2006*, pages 65–68. Fortaleza, Brazil, August 5, 2006.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The Automatic Content Extraction (ACE) Program: Tasks, data, & evaluation. In *Proc. LREC 2004*, pages 837–840. Lisbon, Portugal, 26-28 May 2004.

K. Fundel, R. Küffner, and R. Zimmer. 2007. RELEX: Relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371.

J. Hakenberg, U. Leser, C. Plake, H. Kirsch, and D. Rebholz-Schuhmann. 2005. LLL'05 Challenge: Genic Interaction Extraction – Identification of language patterns based on alignment and finite state automata. *Proc. 4th Learning Language in Logic Workshop*, pages 38–45. Bonn, August 2005.

L. Hirschman, M. Krallinger, and A. Valencia, editors. 2007. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. Madrid: CNIO Centro Nacional de Investigaciones Oncológicas.

M. Huang, X. Zhu, D. Payan, K. Qu, and M. Li. 2004. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

T. Jenssen, A. Lægreid, J. Komorowski, and E. Hovig. 2001. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28:21–28.

N. Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proc. ACL 2004 on Interactive Poster and Demonstration Sessions*, pages 22–25. Barcelona, Spain, July 21-26, 2004.

S. Katrenko and P. Adriaans. 2006. Learning relations from biomedical corpora using dependency trees. *Proceedings of the KDECB 2006 Workshop*, pages 61–80. Ghent, Belgium, May 10, 2006.

C. Nédellec. 2005. Learning Language in Logic: Genic interaction extraction challenge. *Proceedings of the 4th Learning Language in Logic Workshop*, pages 31–37. Bonn, Germany, August 2005.

T. Ohta, Y. Tateisi, and J.-D. Kim. 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain. In *Proceedings of the HLT 2002*, pages 82–86. San Diego, CA, USA, March 24-27, 2002.

S. Pyysalo, F. Ginter, J. Heimonen, J. Björne, J. Boberg, J. Järvinen, and T. Salakoski. 2007. BIOINFER: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(50).

S. Pyysalo, A. Airola, J. Heimonen, J. Björne, F. Ginter, and T. Salakoski. 2008. Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, 9(Supplement 3):S6.

R. Sætre, K. Sagae, and J. Tsujii. 2007. Syntactic features for protein-protein interaction extraction. *Short Paper Proceedings of the LBM 2007*, Singapore, December 6-7, 2007.

J. Saric, L. Jensen, P. Bork, R. Ouzounova, and I. Rojas. 2004. Extracting regulatory gene expression networks from PUBMED. In *Proceedings of the ACL 2004*, pages 191–198. Barcelona, Spain, July 21-26, 2004.

A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii. 2001. Event extraction from biomedical papers using a full parser. *Proceedings of the PSB 2001*, pages 408–419. Maui, Hawaii, USA. January 3-7, 2001.

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

D. Zelenko, C. Aone, and A. Richardella. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research*, 3:1083–1106.

G. Zhou, J. Su, J. Zhang, and M. Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the ACL'05*, pages 427–434. Ann Arbor, MI, USA, 25-30 June 2005.