Extracting Gene Regulation Networks Using Linear-Chain Conditional Random Fields and Rules

Slavko Žitnik[†‡] **Marinka Žitnik**[†] **Blaž Zupan**[†] **Marko Bajec**[†]

[†]Faculty of Computer and Information Science
University of Ljubljana
Tržaška cesta 25
SI-1000 Ljubljana
{name.surname}@fri.uni-lj.si

[‡]Optilab d.o.o.
Dunajska cesta 152
SI-1000 Ljubljana

Abstract

Published literature in molecular genetics may collectively provide much information on gene regulation networks. Dedicated computational approaches are required to sip through large volumes of text and infer gene interactions. We propose a novel sieve-based relation extraction system that uses linear-chain conditional random fields and rules. Also, we introduce a new skip-mention data representation to enable distant relation extraction using first-order models. To account for a variety of relation types, multiple models are inferred. The system was applied to the BioNLP 2013 Gene Regulation Network Shared Task. Our approach was ranked first of five, with a slot error rate of 0.73.

1 Introduction

In recent years we have witnessed an increasing number of studies that use comprehensive PubMed literature as an additional source of information. Millions of biomedical abstracts and thousands of phenotype and gene descriptions reside in online article databases. These represent an enormous amount of knowledge that can be mined with dedicated natural language processing techniques. However, extensive biological insight is often required to develop text mining techniques that can be readily used by biomedical experts. Profiling biomedical research literature was among the first approaches in disease-gene prediction and is now becoming invaluable to researchers (Piro and Di Cunto, 2012; Moreau and Tranchevent, 2012). Information from publication repositories was often merged with other databases. Successful examples of such integration include an OMIM database on human genes and genetic phenotypes (Amberger et al., 2011), GeneRIF function annotation database (Osborne et al., 2006), Gene Ontology (Ashburner et al., 2000) and clinical information about drugs in the DailyMed database (Polen et al., 2008). Biomedical literature mining is a powerful way to identify promising candidate genes for which abundant knowledge might already be available.

Relation extraction (Sarawagi, 2008) can identify semantic relationships between entities from text and is one of the key information extraction tasks. Because of the abundance of publications in molecular biology computational methods are required to convert text into structured data. Early relation extraction systems typically used hand-crafted rules to extract a small set of relation types (Brin, 1999). Later, machine learning methods were adapted to support the task and were trained over a set of predefined relation types. In cases where no tagged data is available, some unsupervised techniques offer the extraction of relation descriptors based on syntactic text properties (Bach and Badaskar, 2007). Current state-of-the-art systems achieve best results by combining both machine learning and rule-based approaches (Xu et al., 2012).

Information on gene interactions are scattered in data resources such as PubMed. The reconstruction of gene regulatory networks is a longstanding but fundamental challenge that can improve our understanding of cellular processes and molecular interactions (Sauka-Spengler and Bronner-Fraser, 2008). In this study we aimed at extracting a gene regulatory network of the popular model organism the *Bacillus subtilis*. Specifically, we focused on the sporulation function, a type of cellular differentiation and a well-studied cellular function in *B. subtilis*.

We describe the method that we used for our participation in the BioNLP 2013 Gene Regulation Network (GRN) Shared Task (Bossy et al., 2013). The goal of the task was to retrieve the genic interactions. The participants were provided with manually annotated sentences from research literature that contain entities, events and genic interactions. Entities are sequences of text that identify objects, such as genes, proteins and regulons. Events and relations are described by type, two associated entities and direction between the two entities. The participants were asked to predict relations of interaction type in the test data set. The submitted network of interactions was compared to the reference network and evaluated with Slot Error Rate (SER) (Makhoul et al., 1999) $SER = (S + I + D)/N$ that measures the fraction of incorrect predictions as the sum of relation substitutions (S), insertions (I) and deletions (D) relative to the number of reference relations (N).

We begin with a description of related work and the background of relation extraction. We then present our extension of linear-chain conditional random fields (CRF) with skip-mentions (Sec. 3). Then we explain our sieve-based system architecture (Sec. 4), which is the complete pipeline of data processing that includes data preparation, linear-chain CRF and rule based relation detection and data cleaning. Finally, we describe the results at BioNLP 2013 GRN Shared Task (Sec. 6).

2 Related Work

The majority of work on relation extraction focuses on binary relations between two entities. Most often, the proposed systems are evaluated against social relations in ACE benchmark data sets (Bunescu and Mooney, 2005; Wang et al., 2006). There the task is to identify pairs of entities and assign them a relation type. A number of machine learning techniques have been used for relation extraction, such as sequence classifiers, including HMM (Freitag and McCallum, 2000), CRF (Lafferty et al., 2001) and MEMM (Kambhatla, 2004), and binary classifiers. The latter most oftem employ SVM (Van Landeghem et al., 2012).

The ACE 2004 data set (Mitchell et al., 2005) contains two-tier hierarchical relation types. Thus, a relation can have another relation as an attribute and second level relation must have only atomic attributes. Therefore, two-tier relation hierarchies have the maximum height of two. Wang et al. (2006) employed a one-against-one SVM classifier to predict relations in ACE 2004 data set using semantic features from WordNet (Miller, 1995). The BioNLP 2013 GRN Shared Task aims to detect three-tier hierarchical relations. These relations describe interactions that can have events or other interactions as attributes. In contrast to pairwise approach of Wang et al. (2006), we extract relations with sequence classifiers and rules.

The same relation in text can be expressed in many forms. Machine-learning approaches can resolve this heterogeneity by training models on large data sets using a large number of feature functions. Text-based features can be constructed through application of feature functions. An approach to overcome low coverage of different relation forms was proposed by Garcia and Gamallo (2011). They introduced a lexico-syntactic pattern-based feature functions that identify dependency heads and extracts relations. Their approach was evaluated over two relation types in two languages and achieved good results. In our study we use rules to account for the heterogeneity of relation representation.

Generally, when trying to solve a relation extraction task, data sets are tagged using the IOB (inside-outside-beginning) notation (Ramshaw and Marcus, 1995), such that the first word of the relation is tagged as *B-REL*, other consecutive words within it as *I-REL* and all others as *O*. The segment of text that best describes a predefined relation between two entities is called a relation descriptor. Li et al. (2011) trained a linear-chain CRF to uncover these descriptors. They also transformed subject and object *mentions* of the relations into dedicated values that enabled them to correctly predict relation direction. Additionally, they represented the whole relation descriptor as a single word to use long-range features with a first-order model. We use a similar model but propose a new way of token sequence transformation which discovers the exact relation and not only the descriptor. Banko and Etzioni (2008) used linear models for the extraction of open relations (i.e. extraction of general relation descriptors without any knowledge about specific target relation type). They first characterized the type of relation appearance in the text according to lexical and syntactic patterns and then trained a CRF using these data along with synonym detection (Yates and Etzioni, 2007). Their method is useful when a few relations in a massive corpus are unknown. However, if higher levels of recall are desired, traditional relation extraction is a better fit. In this study we therefore propose a completely supervised relation extraction method.

Methods for biomedical relation extraction have been tested within several large evaluation initiatives. The Learning language in logic (LLL) challenge on genic interaction extraction (Nédellec, 2005) is similar to the BioNLP 2013 GRN Shared Task, which contains a subset of the LLL data set enriched with additional annotations. Giuliano et al. (2006) solved the task using an SVM classifier with a specialized local and global context kernel. The local kernel uses only mention-related features such as word, lemma and part-of-speech tag, while the global context kernel compares words that appear on the left, between and on the right of two candidate mentions. To detect relations, they select only documents containing at least two mentions and generate $\binom{n}{k}$ training examples, where n is the number of all mentions in a document and k is number of mentions that form a relation (i.e. two). They then predict three class values according to direction (subject-object, object-subject, no relation). Our approach also uses context features and syntactic features of neighbouring tokens. The direction of relations predicted in our model is arbitrary and it is further determined using rules.

The BioNLP 2011 REL Supporting Shared Task addressed the extraction of entity relations. The winning TESS system (Van Landeghem et al., 2012) used SVMs in a pipeline to detect entity nodes, predict relations and perform some post-processing steps. They predict relations among every two mention pairs in a sentence. Their study concluded that the term detection module has a strong impact on the relation extraction module. In our case, protein and entity mentions (i.e. mentions representing genes) had already been identified, and we therefore focused mainly on extraction of events, relations and event modification mentions.

3 Conditional Random Fields with Skip-Mentions

Conditional random fields (CRF) (Lafferty et al., 2001) is a discriminative model that estimates joint distribution $p(\overline{y}|\overline{x})$ over the target sequence $\overline{y}$, conditioned on the observed sequence $\overline{x}$. The following example shows an observed sequence $\overline{x}$ where mentions are printed in bold:

> "Transcription of **cheV** initiates from a **sigma D**-dependent **promoter** element both in vivo and in vitro, and expression of a **cheV**-lacZ fusion is completely dependent on **sigD**." [1]

Corresponding sequences $\overline{x}^{POS}$, $\overline{x}^{PARSE}$, $\overline{x}^{LEMMA}$ contain part-of-speech tags, parse tree tokens and lemmas for each word, respectively. Different feature functions f_j (Fig. 2), employed by CRF, use these sequences in order to model the target sequence $\overline{y}$, which also corresponds to tokens in $\overline{x}$. Feature function modelling is an essential part when training CRF. Selection of feature functions contributes the most to an increase of precision and recall when training CRF classifiers. Usually these are given as templates and the final features are generated by scanning the entire training data set. The feature functions used in our model are described in Sec. 3.1.

CRF training finds a weight vector w that predicts the best possible (i.e. the most probable) sequence $\hat{y}$ given $\overline{x}$. Hence,

$$\hat{y} = \arg\max_{\overline{y}} p(\overline{y}|\overline{x}, w), \qquad (1)$$

where the conditional distribution equals

$$p(\overline{y}|\overline{x}, w) = \frac{\exp(\sum_{j=1}^{m} w_j \sum_{i=1}^{n} f_j(\overline{y}, \overline{x}, i))}{C(\overline{x}, w)}. \qquad (2)$$

Here, n is the length of the observed sequence $\overline{x}$, m is the number of feature functions and $C(\overline{x}, w)$ is a normalization constant computed over all possible $\overline{y}$. We do not consider the normalization constant because we are not interested in exact target sequence probabilities. We select only the target sequence that is ranked first.

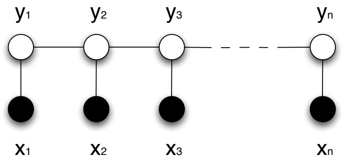

Figure 1: The structure of a linear-chain CRF model. It shows an observable sequence $\overline{x}$ and target sequence $\overline{y}$ containing n tokens.

The structure of a linear-chain CRF (LCRF) model or any other more general graphical model is defined by references to the target sequence labels within the feature functions. Fig. 1 shows the

[1] The sentence is taken from BioNLP 2013 GRN training data set, article PMID-8169223-S5.

```
function f($\overline{y}$, $\overline{x}$, $i$):
    if ($y_{i-1}$ == O and
        $y_i$ == GENE and
        $x_{i-1}$ == transcribes) then
       return 1
    else
       return 0
```

Figure 2: An example of a feature function. It checks if the previous label was *Other*, the current is *Gene* and the previous word was "*transcribes*", returns 1, otherwise 0.

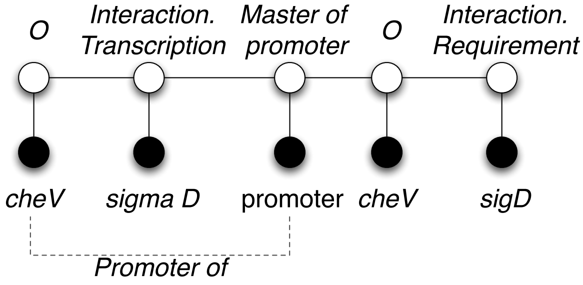

Figure 3: A mention sequence with zero skip-mentions. This continues our example from Sec. 3.

structure of the LCRF. Note that the i-th factor can depend only on the current and the previous sequence labels y_i and y_{i-1}. LCRF can be efficiently trained, whereas exact inference of weights in CRF with arbitrary structure is intractable due to an exponential number of partial sequences. Thus, approximate approaches must be adopted.

3.1 Data Representation

The goal of our task is to identify relations between two selected mentions. If we process the input sequences as is, we cannot model the dependencies between two consecutive mentions because there can be many other tokens in between. From an excerpt of the example in the previous section, "**cheV** initiates from a **sigmaD**", we can observe the limitation of modelling just two consecutive tokens. With this type of labelling it is hard to extract the relationships using a first-order model. Also, we are not interested in identifying relation descriptors (i.e. segments of text that best describe a pre-defined relation); therefore, we generate new sequences containing only mentions. Mentions are also the only tokens that can be an attribute of a relation. In Fig. 3 we show the transformation of our example into a mention sequence. The observable sequence $\overline{x}$ contains sorted entity mentions that are annotated. These annotations were part of the training corpus. The target sequence $\overline{y}$ is tagged with the none symbol (i.e. *O*) or the name of the relationship (e.g. *Interaction.Requirement*). Each relationship target token represents a relationship between the current and the previous observable mention.

The mention sequence as demonstrated in Fig. 3 does not model the relationships that exist between distant mentions. For example, the mentions *cheV* and *promoter* are related by a *Promoter of* relation, which cannot be identified using only LCRF. Linear model can only detect dependencies between two consecutive mentions. To model such relationships on different distances we generate appropriate skip-mention sequences. The notion of *skip-mention* stands for the number of other mentions between two consecutive mentions which are not included in a specific skip-mention sequence. Thus, to model relationships between every second mention, we generate two one skip-mention sequences for each sentence. A one skip-mention sequence identifies the *Promoter of* relation, shown in Fig. 4.

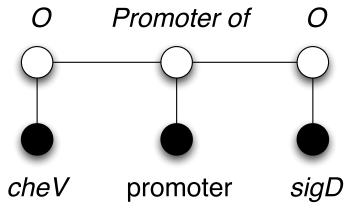

Figure 4: A mention sequence with one skip-mention. This is one out of two generated mention sequences with one skip-mention. The other consists of tokens *sigmaD* and *cheV*.

For every s skip-mention number, we generate $s+1$ mention sequences of length $\lceil \frac{n}{s} \rceil$. After these sequences are generated, we train one LCRF model per each skip-mention number. Model training and inference of predictions can be done in parallel due to the sequence independence. Analogously, we generate model-specific skip-mention sequences for inference and get target labellings as a result. We extract the identified relations between the two mentions and represent them as an undirected graph.

Fig. 5 shows the distribution of distances be-

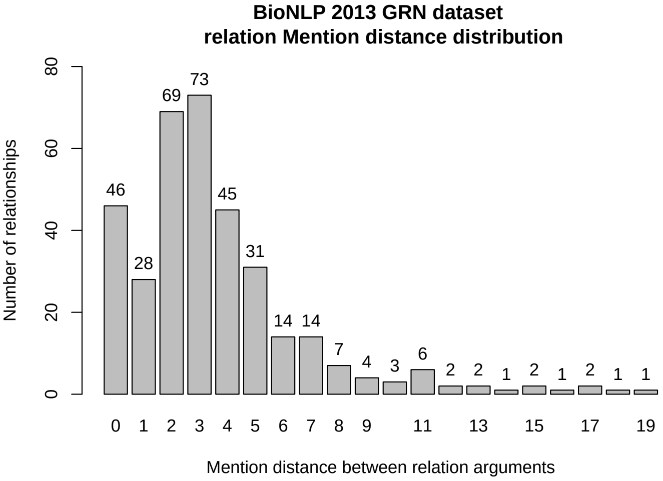

Figure 5: Distribution of distances between two mentions connected with a relation.

tween the relation mention attributes (i.e. agents and targets) in the BioNLP 2013 GRN training and development data set. The attribute mention data consists of all entity mentions and events. We observe that most of relations connect attributes on distances of two and three mentions.

To get our final predictions we train CRF models on zero to ten skip-mention sequences. We use the same unigram and bigram feature function set for all models. These include the following:

- target label distribution,
- mention type (e.g. *Gene*, *Protein*) and observable values (e.g., *sigma D*) of mention distance 4 around current mention,
- context features using bag-of-words matching on the left, between and on the right side of mentions,
- hearst concurrence features (Bansal and Klein, 2012),
- token distance between mentions,
- parse tree depth and path between mentions,
- previous and next lemmas and part-of-speech tags.

4 Data Analysis Pipeline

We propose a pipeline system combining multiple processing sieves. Each sieve is an independent data processing component. The system consists of eight sieves, where the first two sieves prepare data for relation extraction, main sieves consist of linear-chain CRF and rule-based relation detection, and the last sieve cleans the output data. Full implementation is publicly available (https://bitbucket.org/szitnik/iobie). We use CRF-Suite (http://www.chokkan.org/software/crfsuite) for faster CRF training and inference.

First, we transform the input data into a format appropriate for our processing and enrich the data with lemmas, parse trees and part-of-speech tags. We then identify additional action mentions which act as event attributes (see Sec. 4.3). Next, we employ the CRF models to detect events. We treat events as a relation type. The main relation processing sieves detect relations. We designed several processing sieves, which support different relation attribute types and hierarchies. We also employ rules at each step to properly set the agent and target attributes. In the last relation processing sieve, we perform rule-based relation extraction to detect high precision relations and boost the recall. In the last step we clean the extracted results and export the data.

The proposed system sieves are executed in the following order:

i Preprocessing Sieve

ii Mention Processing Sieve

iii Event Processing Sieve

iv Mention Relations Processing Sieve

v Event Relations Processing Sieve

vi Gene Relations Processing Sieve

vii Rule-Based Relations Processing Sieve

viii Data Cleaning Sieve

In the description of the sieves in the following sections, we use general relation terms, naming the relation attributes as subject and object, as shown in Fig. 6.

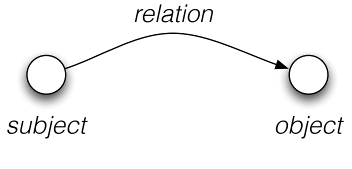

Figure 6: General relation representation.

4.1 Preprocessing Sieve

The preprocessing sieve includes data import, sentence detection and text tokenization. Additionally, we enrich the data using part-of-speech tags, parse trees (http://opennlp.apache.org) and lemmas (Juršic et al., 2010).

4.2 Mention Processing Sieve

The entity mentions consist of *Protein, GeneFamily, ProteinFamily, ProteinComplex, PolymeraseComplex, Gene, Operon, mRNA, Site, Regulon* and *Promoter* types. Action mentions (e.g. *inhibits*, *co-transcribes*) are automatically detected as they are needed as event attributes for the event extraction. We therefore select all lemmas of the action mentions from the training data and detect new mentions from the test data set by comparing lemma values.

4.3 Event Processing Sieve

The general definition of an event is described as a change on the state of a bio-molecule or biomolecules (e.g. "*expression* of a cheV-lacZ fusion is completely dependent on *sigD*"). We represent events as a special case of relationship and name them "*EVENT*". In the training data, the event subject types are *Protein, GeneFamily, PolymeraseComplex, Gene, Operon, mRNA, Site, Regulon* and *Promoter* types, while the objects are always of the action type (e.g. "*expression*"), which we discover in the previous sieve. After identifying event relations using the linear-chain CRF approach, we apply a rule that sets the action mention as an object and the gene as a subject attribute for every extracted event.

4.4 Relations Processing Sieves

According to the task relation properties (i.e. different subject and object types), we extract relations in three phases (iv, v, vi). This enables us to extract hierarchical relations (i.e. relation contains another relation as subject or object) and achieve higher precision. All sieves use the proposed linear-chain CRF-based extraction. The processing sieves use specific relation properties and are executed as follows:

(iv) First, we extract relations that contain only entity mentions as attributes (e.g. "Transcription of cheV initiates from a sigmaD" resolves into the relation *sigmaD* → *Interaction.Transcription* → *cheV*).

(v) In the second stage, we extract relations that contain at least one event as their attribute. Prior to execution we transform events into their mention form. Mentions generated from events consist of two tokens. They are taken from the event attributes and the new *event mention* is included into the list of existing mentions. Its order within the list is determined by the index of the lowest mention token. Next, relations are identified following the same principle as in the first step.

(vi) According to an evaluation peculiarity of the challenge, the goal is to extract possible interactions between genes. Thus, when a relation between a gene $G1$ and an event E should be extracted, the GRN network is the same as if the method identifies a relation between a gene $G1$ and gene $G2$, if $G2$ is the object of event E. We exploit this notion by generating training data to learn relation extraction only between *B. subtilis* genes. During this step we use an external resource of all known genes of the bacteria retrieved from the NCBI[2].

The training and development data sets include seven relation instances that have a relation as an attribute. We omitted this type of hierarchy extraction due to the small number of data instances and execution of relation extraction between genes.

There are also four negative relation instances. The BioNLP task focuses on positive relations, so there would be no increase in performance if negative relations were extracted. Therefore, we extract only positive relations. According to the data set, we could simply add a separate sieve which would extract negations by using manually defined rules. Words that explicitly define these negations are *not*, *whereas*, *neither* and *nor*.

4.5 Rule-Based Relations Processing Sieve

The last step of relation processing uses rules that extract relations with high precision. General rules consist of the following four methods:

- The method that checks all consequent mention triplets that contain exactly one action mention. As input we set the index of the action mention within the triplet, its matching regular expression and target relation.

[2] `http://www.ncbi.nlm.nih.gov/nuccore/AL009126`

- The method that processes every two consequent *B. subtilis* entity mentions. It takes a regular expression, which must match the text between the mentions, and a target relation.

- The third method is a modification of the previous method that supports having a list of entity mentions on the left or the right side. For example, this method extracts two relations in the following example: "*rsfA* is under the control of both *sigma(F)* and *sigma(G)*".

- The last method is a variation of the second method, which removes subsentences between the two mentions prior to relation extraction. For example, the method is able to extract distant relation from the following example: "*sigma(F)* factor turns on about 48 genes, including the gene for *RsfA*, and the gene for *sigma(G)*". This is *sigma(F)* $\rightarrow$ *Interaction.Activation* $\rightarrow$ *sigma(G).*

We extract the Interaction relations using regular expression and specific keywords for the transcription types (e.g. keywords *transcrib*, *directs transcription*, *under control of*), inhibition (keywords *repress*, *inactivate*, *inhibits*, *negatively regulated by*), activation (e.g. keywords *governed by*, *activated by*, *essential to activation*, *turns on*), requirement (e.g. keyword *require*) and binding (e.g. keywords *binds to*, *-binding*). Notice that in biomedical literature, a multitude of expressions are often used to describe the same type of genetic interaction. For instance, researchers might prefer using the expression *to repress* over *to inactivate* or *to inhibit*. Thus, we exploit these synsets to improve the predictive accuracy of the model.

4.6 Data Cleaning Sieve

The last sieve involves data cleaning. This consists of removing relation loops and eliminating redundancy.

A relation is considered a loop if its attribute mentions represent the same entity (i.e. mentions corefer). For instance, sentence "... *sigma D* element, while cheV-lacZ depends on *sigD* ..." contains mentions *sigma D* and *sigD*, which cannot form a relationship because they represent the same gene. By removing loops we reduce the number of insertions. Removal of redundant relations does not affect the final score.

5 Data in BioNLP 2013 GRN Challenge

Table 1 shows statistics of data sets used in our study. For the test data set we do not have tagged data and therefore cannot show the detailed evaluation analysis for each sieve. Each data set consists of sentences extracted from PubMed abstracts on the topic of the gene regulation network of the sporulation of *B. subtilis*. The sentences in both the training and the development data sets are manually annotated with entity mentions, events and relations. Real mentions in Table 1 are the mentions that refer to genes or other structures, while action mentions refer to event attributes (e.g. transcription). Our task is to extract *Interaction* relations of the types *regulation*, *inhibition*, *activation*, *requirement*, *binding* and *transcription* for which the extraction algorithm is also evaluated.

The extraction task in GRN Challenge is twofold: given annotated mentions, a participant needs to identify a relation and then determine the role of relation attributes (i.e. subject or object) within the previously identified relation. Only predictions that match the reference relations by both relation type and its attributes are considered as a match.

6 Results and Discussion

We tested our system on the data from BioNLP 2013 GRN Shared Task using the leave one out cross validation on the training data and achieved a SER of 0.756, with 4 substitutions, 81 deletions, 14 insertions and 46 matches, given 131 reference relations. The relatively high number of deletions in these results might be due to ambiguities in the data. We identified the following number of extracted relations in the relation extraction sieves (Sec. 4): (iii) 91 events, (iv) 130 relations between mentions only, (v) 27 relations between an event and a mention, (vi) 39 relations between entity mentions, and (vii) 44 relations using only rules. Our approach consists of multiple submodules, each designed for a specific relation attribute type (e.g. either both attributes are mentions, or an event and a mention, or both are genes). Also, the total sum of extracted relations exceeds the number of final predicted relations, which is a consequence of their extraction in multiple sieves. Duplicates and loops were removed in the data cleaning sieve.

The challenge test data set contains 290 mentions across 67 sentences. To detect relations

Data set	Documents	Tokens	Real mentions	Action mentions	Events	Relations	Interaction relations
dev	48	1321	205	55	72	105	71
train	86	2380	422	102	157	254	159
test	67	1874	290	86	/	/	/

Table 1: BioNLP 2013 GRN Shared Task development (dev), training (train) and test data set properties.

in the test data, we trained our models on the joint development and training data. At the time of submission we did not use the gene relations processing sieve (see Sec. 4) because it had not yet been implemented. The results of the participants in the challenge are shown in Table 2. According to the official SER measure, our system (U. of Ljubljana) was ranked first. The other four competing systems were K. U. Leuven (Provoost and Moens, 2013), TEES-2.1 (Björne and Salakoski, 2013), IRISA-TexMex (Claveau, 2013) and EVEX (Hakala et al., 2013). Participants aimed at a low number of substitutions, deletions and insertions, while increasing the number of matches. We got the least number of substitutions and fairly good results in the other three indicators, which gave the best final score. Fig. 7 shows the predicted gene regulation network with the relations that our system extracted from test data. This network does not exactly match our submission due to minor algorithm modifications after the submission deadline.

Participant	**S**	**D**	**I**	**M**	**SER**
U. of Ljubljana	**8**	50	6	30	**0.73**
K. U. Leuven	15	53	5	20	0.83
TEES-2.1	9	59	8	20	0.86
IRISA-TexMex	27	**25**	28	**36**	0.91
EVEX	10	67	**4**	11	0.92

Table 2: BioNLP 2013 GRN Shared Task results. The table shows the number of substitutions (S), deletions (D), insertions (I), matches (M) and slot error rate (SER) metric.

7 Conclusion

We have proposed a sieve-based system for relation extraction from text. The system is based on linear-chain conditional random fields (LCRF) and domain-specific rules. In order to support the extraction of relations between distant mentions, we propose an approach called *skip-mention linear chain CRF*, which extends LCRF by varying

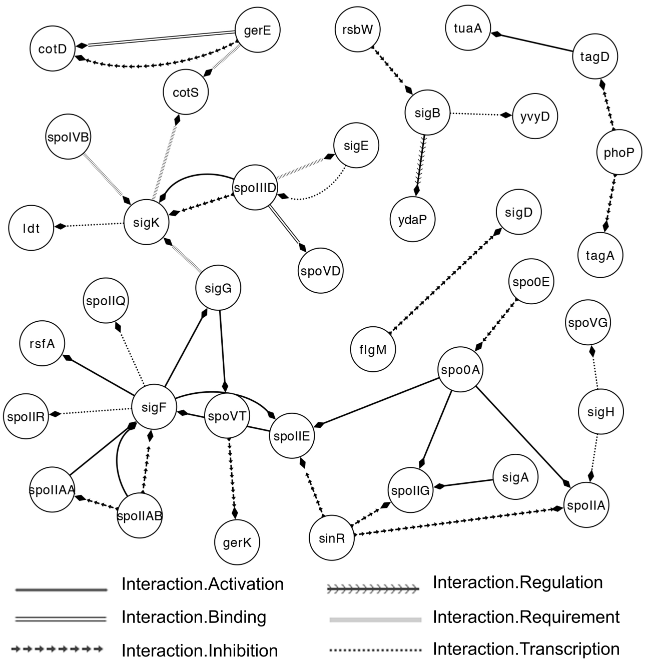

Figure 7: The predicted gene regulation network by our system at the BioNLP 2013 GRN Shared Task.

the number of skipped mentions to form mention sequences. In contrast to common relation extraction approaches, we inferred a separate model for each relation type.

We applied the proposed system to the BioNLP 2013 Gene Regulation Network Shared Task. The task was to reconstruct the gene regulation network of sporulation in the model organism *B. subtilis*. Our approach scored best among this year's submissions.

Acknowledgments

The work has been supported by the Slovene Research Agency ARRS within the research program P2-0359 and in part financed by the European Union, European Social Fund.

References

Joanna Amberger, Carol Bocchini, and Ada Hamosh. 2011. A new face and new challenges for online Mendelian inheritance in man (OMIM). *Human Mutation*, 32(5):564–567.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, Michael J. Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29.

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature Review for Language and Statistics II*, pages 1–15.

Michele Banko and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. *Proceedings of ACL-08: HLT*, page 28–36.

Mohit Bansal and Dan Klein. 2012. Coreference semantics from web features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, page 389–398.

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*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Robert Bossy, Philippe Bessiȓes, and Claire Nédellec. 2013. BioNLP shared task 2013 - an overview of the genic regulation network task. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*, page 172–183. Springer.

Razvan C. Bunescu and Raymond J. Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, page 724–731.

Vincent Claveau. 2013. IRISA participation to bioNLP-ST13: lazy-learning and information retrieval for information extraction tasks. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Dayne Freitag and Andrew McCallum. 2000. Information extraction with HMM structures learned by stochastic optimization. In *Proceedings of the National Conference on Artificial Intelligence*, page 584–589.

Marcos Garcia and Pablo Gamallo. 2011. Dependency-based text compression for semantic relation extraction. *Information Extraction and Knowledge Acquisition*, page 21.

Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. 2006. Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, page 401–408.

Kai Hakala, Sofie Van Landeghem, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Matjaž Juršic, Igor Mozetič, Tomaž Erjavec, and Nada Lavrač. 2010. LemmaGen: multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science*, 16(9):1190–1214.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289.

Yaliang Li, Jing Jiang, Hai L. Chieu, and Kian M.A. Chai. 2011. Extracting relation descriptors with conditional random fields. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 392–400, Thailand. Asian Federation of Natural Language Processing.

John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. 1999. Performance measures for information extraction. In *Proceedings of DARPA Broadcast News Workshop*, page 249–252.

George A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. ACE 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*.

Yves Moreau and Léon-Charles Tranchevent. 2012. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13(8):523–536.

Claire Nédellec. 2005. Learning language in logic-genic interaction extraction challenge. In *Proceedings of the 4th Learning Language in Logic Workshop (LLL05)*, volume 7, pages 1–7.

John D. Osborne, Simon Lin, Warren A. Kibbe, Lihua J. Zhu, Maria I. Danila, and Rex L. Chisholm. 2006. GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM. Technical report, Northwestern University.

Rosario M Piro and Ferdinando Di Cunto. 2012. Computational approaches to disease-gene prediction: rationale, classification and successes. *The FEBS Journal*, 279(5):678–96.

Hyla Polen, Antonia Zapantis, Kevin Clauson, Jennifer Jebrock, and Mark Paris. 2008. Ability of online drug databases to assist in clinical decision-making with infectious disease therapies. *BMC Infectious Diseases*, 8(1):153.

Thomas Provoost and Marie-Francine Moens. 2013. Detecting relations in the gene regulation network. In *Proceedings of BioNLP Shared Task 2013 Workshop*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Lance A. Ramshaw and Mitchell P. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*, page 82–94.

Sunita Sarawagi. 2008. Information extraction. *Foundations and Trends in Databases*, 1(3):261–377.

Tatjana Sauka-Spengler and Marianne Bronner-Fraser. 2008. A gene regulatory network orchestrates neural crest formation. *Nature reviews Molecular cell biology*, 9(7):557–568.

Sofie Van Landeghem, Jari Björne, Thomas Abeel, Bernard De Baets, Tapio Salakoski, and Yves Van de Peer. 2012. Semantically linking molecular entities in literature through entity relationships. *BMC Bioinformatics*, 13(Suppl 11):S6.

Ting Wang, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham, and Ji Wang. 2006. Automatic extraction of hierarchical relations from text. *The Semantic Web: Research and Applications*, page 215–229.

Yan Xu, Kai Hong, Junichi Tsujii, I Eric, and Chao Chang. 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *Journal of the American Medical Informatics Association*, 19(5):824–832.

Alexander Yates and Oren Etzioni. 2007. Unsupervised resolution of objects and relations on the web. In *Proceedings of NAACL HLT*, page 121–130.