Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016

Estelle Chaix[*], **Bertrand Dubreucq**[$], **Abdelhak Fatihi**[$], **Dialekti Valsamou**[*,&],
Robert Bossy[*], Mouhamadou Ba[*], Louise Deléger[*] , Pierre Zweigenbaum[&],
Philippe Bessières[*], Loic Lepiniec[$], Claire Nédellec[*]
[*] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[$] IJPB, INRA, AgroParisTech, CNRS, Université Paris-Saclay, 78026 Versailles, France
[&] LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
[*]forename.lastname@jouy.inra.fr
[$]forename.lastname@versailles.inra.fr;[&] pz@limsi.fr

Abstract

This paper presents the SeeDev Task of the BioNLP Shared Task 2016. The purpose of the SeeDev Task is the extraction from scientific articles of the descriptions of genetic and molecular mechanisms involved in seed development of the model plant, *Arabidopsis thaliana*. The SeeDev task consists in the extraction of many different event types that involve a wide range of entity types so that they accurately reflect the complexity of the biological mechanisms. The corpus is composed of paragraphs selected from the full-texts of relevant scientific articles. In this paper, we describe the organization of the SeeDev task, the corpus characteristics, and the metrics used for the evaluation of participant systems. We analyze and discuss the final results of the seven participant systems to the test. The best F-score is 0.432, which is similar to the scores achieved in similar tasks on molecular biology.

1 Introduction

Since its first edition in 2009, BioNLP Shared Task (BioNLP-ST) organizes information extraction (IE) tasks from scientific literature with a focus on molecular mechanisms with the aim to promote advances in IE research in the biomedical domain. The SeeDev task is the first task on event extraction about molecular biology of plants. It gives an opportunity for the BioNLP community to evaluate the reusability of methods, to characterize the peculiarities of IE for the plant biology domain and to develop dedicated approaches. For this purpose, we manually annotated a new corpus of scientific papers selected for their relevance to the topic. We propose to the participants to extract text-bound events that involve biological entities provided as input. The performances of the systems are evaluated by standard measures through the comparison of their predictions to the reference annotations.

2 Context

Seeds are the main vectors for breeding and production of annual field crops. The accumulation of seed storage compounds (*e.g.* sugars, lipids, proteins) is of primary importance for food, feed and industrial uses. Seed development requires the coordinated growth of different tissues that involves complex genetics and environmental regulations (Alberts et al., 2002). A comprehensive understanding of the molecular networks that underlie the regulation of seed development remains a major scientific challenge with important potential impact on fundamental research, agriculture and industry.

The SeeDev task of BioNLP Shared Task 2016 focuses on the accumulation of reserves in the seed of the model plant, *Arabidopsis thaliana* (*Ath*), for which research on regulatory networks is the subject of a large and active international community (Santos-Mendoza et al., 2008). Most of this knowledge is spread in thousands of articles. As such, this topic constitutes an excellent primer for the development of event extraction methods. The SeeDev corpus should then be largely reusable for the study of other plants and other development phases.

Information Extraction research applied to biology mainly consists in automatic entity extraction, their normalization and event extraction (Ananiadou et al., 2014). The extraction of regulatory network has become one of the most popular tasks in shared tasks in recent years. The increasing

Proceedings of the 4th BioNLP Shared Task Workshop, pages 1–11,
Berlin, Germany, August 13, 2016.

complexity of the event scheme over the years is driven by the significant scientific advances in IE and the increasing need for computational models in bioinformatics and systems biology. In 2005, the objective of the *Learning Language in Logic* challenge (LLL'05) was the extraction of gene interactions between proteins and genes with the goal of reconstructing bacterial regulatory networks (Nédellec, 2005). The diversity of the biological events (molecular, physiological) and entities (genes, proteins, families, sites, environmental factors and phenotypes) has continuously increased over the time together with the variety of the biological mechanisms studied. These mechanisms range from detailed networks as in *Bacteria Interaction* (Bossy et al., 2012) and *Gene Regulation Network* (Bossy et al., 2015) tasks, signaling pathways as in *GENIA* task (Kim et al., 2013a) and metabolism to diseases as in *Pathway Curation* (*PC*) and *Cancer Genetics* (*CG*) tasks (Pyysalo et al., 2015). Their extraction from text makes an increasing use of existing standards, nomenclatures and ontologies such as Gene Ontology that facilitates the integration of the text mining results into larger knowledge bases and bioinformatics applications (*e.g.* GRO task (Kim et al., 2013b)) or OntoBiotope (*e.g. Bacteria Biotope* task (Bossy et al., 2015)).

The SeeDev task brings a new application domain, plant development biology, with similar goals and representation as previous IE shared tasks on biological event extraction. This new application domain has required the design of a new knowledge model for the representation of the events, a manually annotated corpus and new metric that accounts for the varying importance of the event arguments.

We refer to the SeeDev task knowledge model as *Gene Regulatory Network for Arabidopsis* (GRNA). GRNA meets the usual constraints of manual annotation of texts (*e.g.* biological relevance and computational tractability), and of automatic annotation by IE methods (*e.g.* learnability from training examples). We have also taken into account the expected use of GRNA for the indexing and retrieval of textual events and experimental data in a unified representation, the modeling of other plant systems, and also the integration of text knowledge with knowledge derived from experimental data.

SeeDev corpus is composed of paragraphs from a selection of recent full-text scientific papers about molecular biology of seed development.

3 Task Description

The SeeDev Task consists in two subtasks (1) *SeeDev-binary* on binary relation extraction and (2) *SeeDev-full* on full event extraction. The *SeeDev-binary* subtask has been conceived as a first step towards the extraction of full n-ary events, which is of interest for plant biology. Both subtasks share the same GRNA model and the same document set with different annotation sets. The two annotations sets contain binary relations and events respectively. The annotation set of *SeeDev-binary* has been computed from the annotation set of *SeeDev-full* through the application of formal transformation rules.

3.1 Knowledge Representation

The GRNA model defines 16 entity types (Figure 1) and 21 event types (Table 1). They are classified into categories and subcategories for readability purpose.

***Molecule**:*
DNA: Gene, Gene_Family, Box, Promoter
DNA product : RNA, Protein, Protein_Family, Protein_Complex, Protein_Domain
Hormone: Hormone
***Dynamic Process**: Regulatory_Network, Pathway*
***Context**: Tissue, Development_Phase, Genotype, Environmental_Factor*

Figure 1. SeeDev entity types.

The *Molecule* category includes molecules that are directly involved in regulation, such as *Hormone* that plays a critical role in plant growth, and *Protein Domain* and DNA regions (*Box*, *Promoter*) for the representation of physical binding events. Protein and gene families are also important entities because they are mentioned as actors of the regulations in some papers without more precision on the exact molecule. The *Dynamic Process* category is defined by two broad entity types, *Regulatory Network* and *Metabolic pathway*, with the purpose of keeping the complexity of the extraction task tractable. Moreover, the distinction in the SeeDev corpus between specific kinds of networks or pathways would have been difficult, if not impossible because the authors themselves remain vague.

Relation Name	**Definition**	**#**	**Train**	**Dev**	**Test**	**Total**
Regulation		1731	46%	22%	31%	48%
Regulates Accumulation (Regulation Of Accumulation)	A Molecule, Dynamic Process or Context regulates the accumulation of a Functional Molecule (in particular, [*Protein*], [*RNA*], [*Hormone*]).	81	44%	36%	20%	2%
Regulates Development Phase (Regulation Of Development Phase)	A Molecule, Dynamic Process or Context regulates the activity of a Development phase.	242	44%	24%	32%	7%
Regulates Expression (Regulation Of Expression)	**A Molecule, Dynamic Process or Context regulates the expression of a DNA entity. DNA entity includes [*Promoter*] and [*Box*].**	**450**	**45%**	**25%**	**31%**	**13%**
Regulates Molecule Activity (Regulation Of Molecule Activity)	An Agent (Molecule, Dynamic Process or Context) regulates the activity of a Molecule, such as [*Protein*].	25	64%	0%	36%	1%
Regulates Process (Regulation Of Process)	**A Molecule, Dynamic Process or Context regulates the activity of a Dynamic Process.**	**904**	**48%**	**20%**	**32%**	**25%**
Regulates Tissue Development (Regulation Of Tissue Development)	A Molecule, Dynamic Process or Context regulates the activity of a Tissue Development.	29	31%	31%	38%	1%
Function		257	42%	28%	30%	7%
Is Involved In Process (Involvement In Process)	A Molecule is involved *in* a Dynamic Process.	55	42%	36%	22%	2%
Transcribes Or Translates To (Transcription Or Translation)	A DNA entity encodes for a RNA (Transcription) or a RNA entity encodes a Protein (Translation). Often, reference is made to the gene encoding the protein, without mention of the RNA.	54	46%	24%	30%	2%
Is Functionally Equivalent To* (Functional Equivalence)	A Molecule, Dynamic Process or Context is compared to a similar entity.	148	41%	26%	33%	4%
Interaction		264	46%	21%	33%	7%
Interacts With (Interaction)	A molecule interacts with another molecule.	148	42%	22%	36%	4%
Binds To (Binding)	A functional molecule physically binds to a molecule.	116	52%	21%	28%	3%
Where and When		704	45%	23%	32%	20%
Exists At Stage (Presence At Stage)	A Molecule is present *during* a Developmental phase.	33	45%	24%	30%	1%
Exists In Genotype (Presence In Genotype)	**A Molecule or Element is present *in* a Genotype**	**377**	**45%**	**21%**	**34%**	**11%**
Occurs During (Occurrence During)	A Process occurs *during* a Developmental Phase.	30	27%	33%	40%	1%
Occurs In Genotype (Occurrence In Genotype)	A Process occurs *in* a Genotype	48	38%	33%	29%	1%
Is Localized In (Localization)	A Molecule is found in a Tissue	216	50%	22%	29%	6%
Composition and Membership		532	44%	22%	34%	15%
Composes Primary Structure (Primary Structure Composition)	A specific sequence of nucleotide is found in a DNA entity.	51	39%	29%	31%	1%
Composes Protein Complex (Protein Complex Description)	A specific DNA product is found in a Protein complex.	19	84%	0%	16%	1%
Has Sequence Identical To* (Sequence Identity)	A Molecule, Dynamic Process or Context is compared to a similar Molecule, Dynamic Process or Context.	126	49%	16%	35%	4%
Is Member Of Family (Family Membership)	A DNA, RNA or Protein belongs to another DNA, Product or Factor. Used between entities of the same nature to denote members of a set.	230	39%	24%	37%	6%
Is Protein Domain Of (Protein Domain Composition)	A specific Protein Domain is found in an amino acid sequence.	106	43%	27%	29%	3%
Specific to Binary scheme		87	51%	26%	23%	2%
Is Linked To*	Used to derive binary relations from n-ary events: it relates optional and main arguments of n-ary events.	87	51%	26%	23%	2%
Total		3575	46%	23%	32%	100%

Table 1: Definition of relations and example distribution in SeeDev *Binary* subtask. Event names are into brackets. (Event arguments are ordered, except events marked with *.)

N-ary representation : Binding							
	Mandatory arguments		Optional arguments				
Role	Functional Molecule	Molecule	Tissue	Developmental Stage	Organism Genotype	Environmental Factor	Hormone
Signature	*RNA, Protein, Protein Family, Protein Complex, Protein Domain, Hormone*	*Gene, Gene Family, Box, Promoter, RNA, Protein, Protein Family, Protein Complex, Protein Domain,*	*Tissue*	*Development Phase*	*Genotype*	*Environmental Factor*	*Hormone*
Binary representation : Binds_to							

Figure 2: Representation of *Binds_to* and *Binding* relation, with mandatory and optional arguments.

Arg 2 Arg 1	Gene	Gene Family	Box	Promoter	RNA	Protein	Protein Family	Protein Complex	Protein Domain	Hormone	Regulatory Network	Metabolic pathway	Genotype	Tissue	Development Phase	Environmental Factor
Gene	5	6	3	3	3	5	5	5	3	3	3	3	1	1	1	1
Gene Family	5	6	3	3	3	5	5	5	3	3	3	3	1	1	1	1
Box	3	3	6	4	2	4	4	4	2	3	3	3	1	1	1	1
Promoter	3	3	4	6	2	4	4	4	2	3	3	3	1	1	1	1
RNA	3	3	3	3	6	6	6	6	4	4	3	3	1	2	2	1
Protein	4	4	4	4	4	7	8	6	3	4	3	3	1	2	2	1
Protein Family	4	4	4	4	4	7	8	6	3	4	3	3	1	2	2	1
Protein Complex	4	4	4	4	4	5	5	8	3	4	3	3	1	2	2	1
Protein Domain	4	4	4	4	4	6	6	7	6	4	3	3	1	2	2	1
Hormone	3	3	3	3	3	4	4	4	2	6	3	3	1	2	2	1
Regulatory Network	2	2	2	2	2	3	3	3	1	3	4	2	1	2	2	1
Metabolic pathway	2	2	2	2	2	3	3	3	1	3	2	4	1	2	2	1
Genotype	1	1	1	1	1	2	2	2	0	2	1	1	3	1	1	0
Tissue	1	1	1	1	1	2	2	2	0	2	1	1	1	3	1	0
Development Phase	1	1	1	1	1	2	2	2	0	2	1	1	1	1	3	0
Environmental Factor	2	2	2	2	2	3	3	3	1	3	2	2	1	2	2	3

Figure 3: Number of relation type by pairs of argument types.

The conditions in which the regulations occur represent critical information about the event context. The entity types represent spatial conditions (*Tissue*), temporal conditions (*Development phase*), the organism, which is genetically modified or not (*Genotype*), and the environmental factors (biotic and abiotic external conditions). The entities in the corpus are denoted by individual words or by sets of words that may be discontinuous.

The 21 GRNA event types are grouped in 6 sets, according to their biological role (Table 1). The *Regulation*, *Function* and *Interaction* categories are central for the description of the biological mechanisms. *Where and When* event types represent the context of the mechanisms, whilst *Composition and Membership* events allow to finely represent relations among the biological entities. Some of the event types, *e.g. Regulates Expression / Process / Molecule Activity* are very similar to those of other molecular biology IE event schemes such as the ones of *GENIA* (Kim et al., 2013a), *Cancer Genetics* (Pyysalo et al., 2015) and *Arabidopsis Leaf Growth* (*LG*) (Szakonyi et al., 2015). Other GRNA event types are specific to biological development, *e.g. Regulates Development Phase / Tissue Development* or to the storage process, *e.g. Regulates Accumulation*. The

LG model of Szakonyi et al. (2015) dedicated to *Ath* does not include plant or development specific events to be reused in GRNA. Protein modification and metabolism in GENIA and PC tasks and regulation of phenotype in *LG*, were not relevant for the SeeDev corpus but will be addressed in priority in further extensions of GRNA.

The first column of Table 1 displays the binary relation names of *SeeDev-binary* subtask and the n-ary event names of *SeeDev-full* subtask in brackets, with their definition in column two. N-ary events have two mandatory arguments and up to five optional arguments: *Tissue*, *Developmental Stage*, *Organism*, *Genotype*, *Environmental Factor*, and *Hormone*.

Furthermore, n-ary events may have a negation modality. Participants are provided with text documents, gold entity annotations, and the detailed signatures of each event, *i.e.* the list of allowed types per slot. Figure 2 gives, for example, the *Binding* event signature.

The use of a strongly typed model facilitates the event prediction because it drastically reduces the number of event candidates given the types of the arguments. Figure 3 shows the number of relation types per pair of argument types. For example the argument pair (*Arg1: Development_Phase* / *Arg2: Protein_Domain*) does not accept any relation type; whereas the pair (*Arg1: Protein* / *Arg2: Protein_Family*) may be involved into 8 different relations. The formal specification of event signatures drastically reduces the exploration space of possible events.

3.2 Sub-Task 1: SeeDev Binary Relation Extraction

The goal of *SeeDev-binary* is the extraction of binary relations of 22 different types without modality (no negation) as described in Table 1. The *Is_Linked_To* relation is computed from the n-ary events, it links mandatory arguments to optional arguments. Figure 4.a gives an example of *SeeDev-binary* annotation with 3 different relations.

3.3 Sub-Task 2: SeeDev Full Relation Extraction

SeeDev-full aims at extracting n-ary events where the number of arguments ranges from two to eight, plus a negation modality. There are three arguments in average. There is no trigger word in SeeDev event representation. Events relate either entities or other events. Figure 4.b gives an example of a *Binding* event with a *Genotype* argument. In the binary version (Figure 4.a), the *Genotype* becomes a mandatory argument of one of the *Exists_In_Genotype* relations.

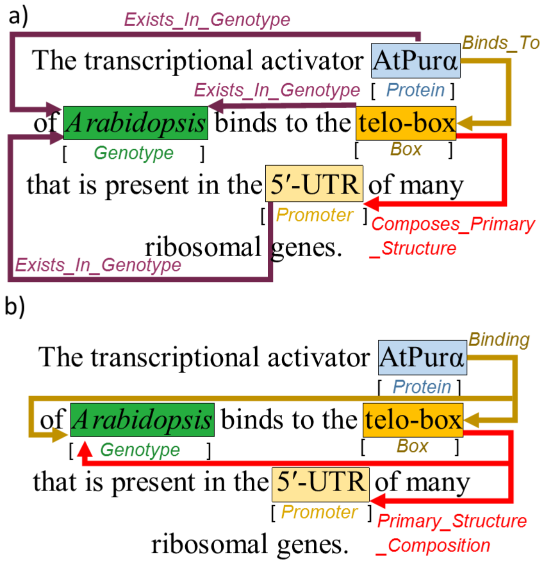
a)
Exists_In_Genotype
Binds_To
The transcriptional activator AtPurα
Exists_In_Genotype
[Protein]
of Arabidopsis binds to the telo-box
[Genotype]
[Box]
that is present in the 5′-UTR of many
[Promoter]
Composes_Primary _Structure
ribosomal genes.
Exists_In_Genotype
b)
Binding
The transcriptional activator AtPurα
[Protein]
of Arabidopsis binds to the telo-box
[Genotype]
[Box]
that is present in the 5′-UTR of many
[Promoter]
Primary_Structure _Composition
ribosomal genes.

Figure 4: Examples of an annotated sentence in (a) *SeeDev-binary* task and (b) *SeeDev-full* task

4 Corpus Description

The SeeDev corpus is a set of 86 paragraphs from 20 full-text articles, selected by plant biology experts, about seed development in *Arabidopsis thaliana*. Table 2 summarizes the SeeDev corpus statistics and data distribution in the Training, Development and Test sets.

	#	Train	Dev	Test
Documents	20	90%	75%	80%
Paragraphs	87	45%	22%	33%
Words	44,857	45%	23%	33%
Entities	7,082	46%	23%	31%
Events	2,583	45%	23%	32%
Relations	3,575	46%	23%	32%

Table 2. SeeDev corpus statistics.

Paragraphs of the same document may be distributed into different sets. The “Documents” row indicates the proportion of documents represented in the set. The SeeDev corpus is smaller than other BioNLP-ST corpora, *e.g.* a fifth of *Cancer Genetics* corpus and a third of *GENIA* corpus. The manual annotation of the SeeDev corpus required a high level of expertise that do not allow for a large corpus, as in many specific domains of Life

Science. We identify small dataset processing as a challenge to overcome by information extraction tools.

Table 1 details the distribution of instances per relation type in the training, development and test sets of the *SeeDev-binary* task. The distribution was balanced between the three data sets so that the test set would represent approximately a third of the annotations for each group of relations. The most frequent relations are *Regulation* with 48% of annotations, which corresponds to what is expected given the corpus domain. The three relations *Regulate Expression*, *Regulates Process* and *Exist in Genotype,* highlighted in Table 1, account for half of the total, whilst seven of the relations are relatively infrequent with 1% of the total.

5 Annotation Methodology

We have successively refined the annotation scheme of GRNA during the annotation process. We have defined an initial annotation scheme according to our expertise in *A. thaliana* seed development and in BioNLP task definition, starting from the GRN model (Bossy et al., 2015).

The scheme was improved through several iterations of manual annotations and collective discussions until it met the requirements, *i.e.* it allowed unambiguous, consistent, readable and detailed formal annotations. Together with the scheme, a very precise guideline document (Chaix et al., 2016) was produced that details the annotation principles for each entity and event type, and provides many examples and counter-examples.

The relevant paragraphs of the corpus were chosen by the biologists, mostly from the abstract, introduction, result and discussion sections. A team of three experts in seed development and two bioinformaticians has manually annotated the corpus following the guidelines by using the AlvisAE Annotation Editor (Papazian et al., 2012) in accordance with the final version of the scheme.

5.1 Automatic Annotation

Rigid designators of named entities, such as *Gene*, *Protein*, *Tissues*, and *Developmental Phases* were automatically pre-annotated with the AlvisNLP pipeline using relevant *Ath* databases (*e.g.* TAIR[1]) and customized lexicons. The goal of automatic pre-annotation was to speed-up the manual annotation process. The evaluation of the automatic annotation compared to the gold standard annotation shows a F-score equal to 0.41, with a high precision (0.89) and low recall (0.26) due to a lack of relevant lexicon for most entity types.

5.2 Manual Annotation

The manual annotation has been achieved in four successive phases in order to both save expert time and achieve a high quality annotation. First, a bioinformatician who is not a specialist of *Ath* annotated all the entities of the corpus. The evaluation of the manual annotation of the entities compared to the gold standard annotation yielded a high 0.93 F-score with balanced Recall and Precision, 0.93 and 0.95 respectively.

Then *Ath* experts revised the entity annotations and annotated the events of the corpus in a double-blind manner. Thanks to the manual pre-annotation of entities, they could focus on events which require more expertise. Next, the annotators together with the bioinformatician used the AlvisAE conflict resolution functionality to build a consensus. Finally, the bioinformatician carefully checked the compliance of each annotation to the guidelines to produce the gold annotation set.

To evaluate the inter-annotator agreement, we measured the F-score between the annotation set of each annotator (referred to as A and B) and the consensus annotation set (*i.e.* gold annotations) (Table 3). The differences between the individual annotators vary according to the event types. The recall measure of the annotations of events with arguments of Process type without regulation (*Is Involved In Process*) and events with Genotype arguments (*Exists In Genotype, Occurs In Genotype*) is lower.

Mistyping *Regulates Accumulation* was frequent because this event is easily confused with *Regulates Molecule Activity*. Annotations from annotator B are closer to the reference annotation, but the examination of the union of both annotation sets shows that annotator B missed events that were well annotated by A. The 0.724 F-score of the union of A and B annotation sets is quite high. The last step of the SeeDev corpus construction is the adjudication between the two annotators with a third person as external referee. It was an essential step to avoid event oversight.

[1] The Arabidopsis Information Resource `http://arabidopsis.org/`

Annotator	F1	Recall	Precision
A	0.548	0.417	0.798
A (T)	+0.048	+0.031	+0.058
B	0.653	0.575	0.754
B (T)	+0.069	+0.071	+0.080
A U B	0.724	0.720	0.728
A U B (T)	**+0.045**	**+0.045**	**+0.045**

Table 3: Evaluation of the inter-annotator agreement by comparing each annotator output to the reference annotation. (T) indicates the gain if relation types are ignored. A U B denotes the union of annotations from annotators A and B.

6 Evaluation Procedure

6.1 Shared Task Organization

As for previous challenges, BioNLP-ST 2016 provides resources and information to the participants through the BioNLP-ST website[2]and mailing lists. The schedule of the SeeDev task follows the usual principles of BioNLP-ST tasks, it can be found on dedicated pages.

We provided state-of-art automatic NLP analysis as supporting resources with the purpose to speed-up the participant system development. Nine tools were selected and applied to the training, development and test sets: POS tagger (*GENIA Tagger* (Tsuruoka et al., 2005)), parsers (*Stanford Parser* (Manning, 2003) *Enju* (Miyao and Tsujii, 2008) *C&C CCG Parser* (Clark and Curran, 2007)), term extractor (*BioYaTeA* (Golik et al., 2013)) named entity recognizers (*Stanford NER* (Finkel et al., 2005) *LINNAEUS* (Gerner et al., 2010) *SR4GN* (Wei et al., 2012)) and tokenizer and sentence splitter (*AlvisNLP suite* (Ba and Bossy, 2016)).

Community web tools (forum, FAQ and mailing list) have been made available on the website with the purpose to federate the community that participates to the challenge. In this way participants could interact with the task organizers and with other participants.

Furthermore, participants could evaluate their predictions through an online evaluation service. During the training phase it was restricted to the evaluation on training and development sets. The service allows now to evaluate predictions on the test set and will remain open. For the first time in BioNLP-ST, participants could also keep track of the performance of various experiments through the same online service. Thus, participants could follow and compare their results and competing team results. The recorded submissions were kept anonymous to other participants. The aim of this tool was to ease the interpretation of the scores and to assist participants in the development-test cycles.

6.2 Evaluation Metrics

The evaluation measures of the participant system results are computed through the comparison of predicted events against reference corpus events. In *SeeDev-binary* the participants had to predict relations between entities given as input. This task can be viewed as a classification task of all pairs of entities. Thus, we evaluate submissions with Recall, Precision and F-score. Submissions were ranked by F-score, however we also provided alternate evaluations in order to assess the strengths of each submission for each relation type separately, for each broad category of relations separately and without taking into account the relation types.

We also designed a measure for *SeeDev-full* task evaluation that is permissive for optional arguments. The evaluation is detailed on the task web site and is available through the online evaluation service to the benefit of teams that will bravely tackle this task.

7 Results

7.1 Participating Systems

Seven teams from 4 continents submitted their results to the test of the SeeDev binary task that are: *DUTIR* (Dalian University of Technology, China), *LIMSI* (CNRS, France), *LitWay* (Xidian University, China), *ULisboa* (LaSIGE, Universidade de Lisboa, Portugal), *UniMelb* (University of Melbourne, Australia), *VERSE* (University of British Columbia, Canada) and *UTS* (University of Turku, Finland).

Their main background domains are Bioinformatics, Machine Learning, Natural Language Processing and Biology according to their responses to a survey.

Table 4 summarizes the scores obtained by the participant systems ranked by F1-score (detailed results are available on the SeeDev site). The results of the *DUTIR* system are not displayed because they experienced a last minute hitch

[2] BioNLP-ST website http://2016.bionlp-st.org/tasks/seedev

and ranked last. *LitWay* from Xidian University achieves the best F1-score (0.432), 0.068 points higher than the second team and 0.177 points higher than the lowest score at 0.255. The two systems that ranked first achieved a balanced recall and precision, while the four others favored recall over precision (*VERSE*, *LIMSI*), or the reverse (*UTS*, *ULISBOA*). *VERSE* obtained the best recall and *UTS* the best precision.

Participant	F1	Recall	Precision
LitWay	0.432	0.448	0.417
UniMelb	0.364	0.386	0.345
VERSE	0.342	0.458	0.273
UTS	0.335	0.245	0.533
ULISBOA	0.306	0.256	0.379
LIMSI	0.255	0.318	0.212

Table 4: Evaluation scores of the SeeDev binary task ranked by F- score.

The best F1-scores are very similar to the ones achieved by participants of previous shared tasks on regulation event extraction around 50% (*e.g.* GRN, CG, PC), which is over what could be expected given the complexity and the novelty of the task and the variability of the example distribution among the events.

As shown by Table 5, the detailed scores per relation exhibit a high variability. Some relations were difficult to predict (*e.g. Regulates Tissue Development*, *Regulates Molecule Activity*, *Occurs During*) while others were well-predicted (*e.g. Composes Primary Structure* with a maximum F1-score of 0.67).

As usual in such corpus, the analysis of the results shows that the causes are multifactorial, we hypothesize that the number of training examples combined with the regularity of the descriptions and the constraints imposed by the event signature are critical. For instance, the *Composes Primary Structure* relation has only 51 examples, but it links entities from a restricted range of types, which makes it easier to predict (0.67 best F1-score). However, other relations such as *Regulates Expression with* a high number of examples (450 examples), inter sentence occurrences (23) and a wide range of argument types (4 types for the first argument and 16 for the second) were poorly predicted (0.39 best F1-score).

The scores of most of the systems remain unchanged when the dataset is restricted to the relations that occur in a single sentence. The difference of the results obtained for intra-sentence dataset are less than 1 point, except for *Limsi* that gains 0.056 points; indeed, *Limsi* is the only team that attempts to predict inter-sentence relations whereas all other participant systems predicted only intra-sentence relations. Given the proportion of inter-sentence relations in the test set (4%), the penalty of ignoring them could have been considered as bearable.

Relation	**Best F1 score**	**System**
All Relations	**0.432**	**LitWay**
Where and When	**0.142**	**LitWay**
Exists_At_Stage	0.167	ULISBOA
Exists_In_Genotype	0.492	LitWay
Occurs_During	0	-
Occurs_In_Genotype	0.167	VERSE
Is_Localized_In	0.450	LitWay
Function	**0.255**	**ULISBOA**
Is_Involved_In_Process	0	-
Transcribes_Or_Translates_To	0.343	VERSE
Is_Functionally_Equivalent_To	0.708	LitWay
Regulation	**0.416**	**LitWay**
Regulates_Accumulation	0.316	UniMelb
Regulates_Development_Phase	0.376	UniMelb
Regulates_Expression	0.386	UniMelb
Regulates_Molecule_Activity	0	
Regulates_Process	0.504	LitWay
Regulates_Tissue_Development	0	-
Composition_MemberShip	**0.490**	**LitWay**
Composes_Primary_Structure	0.667	LIMSI
Composes_Protein_Complex	0.500	UTS
Has_Sequence_Identical_To	0.867	LitWay
Is_Member_Of_Family	0.534	LitWay
Is_Protein_Domain_Of	0.438	LitWay
Interaction	**0.303**	**UniMelb**
Interacts_With	0.286	UniMelb
Binds_To	0.310	VERSE
Is_Linked_To	0.154	VERSE

Table 5: Best F1-score per relation and per category of relation.

In order to assess the difficulty to predict the correct relation type, we computed the F-scores when considering the category of the relations instead of the actual type (first line per category in bold and italic in Table 5). This did not yield a significant improvement although some participants were able to successfully predict events in categories with high biological relevance, such as the *Regulation* category (*Litway* F1: 0.416) and the *Interaction* category (*UniMel* F1: 0.303).

7.2 Systems Description and Result Discussion

All teams used supervised machine-learning approaches (Table 6). Five systems used support vector machines (SVM) and two systems were based on different algorithms, namely *maximum entropy* (MaxEnt) (*LIMSI*) and *convolutional neural network* (*DUTIR*).

Participant	General method
LitWay	Hand crafted patterns + SVM
UniMelb	SVM + Bayes classifiers
VERSE	Linear SVM
ULISBOA	SVM kernel based
UTS	SVM multi-classification
LIMSI	Bag of words
DUTIR	Convolutional neural network

Table 6: General methods of the participants

SVM are widely used for information extraction tasks, because they are powerful versatile classifiers. SVM are kernel-based and there are several existing kernels available (Zelenko et al., 2003) adapted to different object representations. For instance, dependency-path kernels (Bunescu and Mooney, 2005; Airola et al., 2008) handle candidates represented as syntactic dependency paths. Moreover, the usual feature selection methods can be handled by kernels that work on vectorial representations. MaxEnt and neural networks are also popular algorithms in information extraction tasks (McCallum et al., 2000). The most notable characteristic of the best performing system, *LitWay*, is that it combines supervised machine learning for the prediction of a selection of event types with hand-crafted rules for the prediction of other types.

All teams used token segmentation, sentence splitting and token normalization (stemming, lemmatization, POS-tagging). Four teams, among which the three top ranking also used deep syntactic parsing, which confirms that parsing is a powerful pre-processing step for information extraction. Finally, the *LitWay* system also designed features based on word embedding which is a novelty in the BioNLP-ST.

8 Conclusion

We have described the SeeDev task that we have designed with the goal to promote progress in information extraction in the field of plant development and more precisely plant regulatory networks. Two sub-tasks were proposed with increasing levels of complexity, *SeeDev-binary* on binary relations and *SeeDev-full* on events.

The lack of participation to *SeeDev-full* shows that the extraction of n-ary events with optional arguments remains challenging.

Seven teams from different countries participated in the *SeeDev-binary* task with different approaches. The results are very promising, given the novelty of the task and the complexity of the model. The best F-score, 0.432, is close to what has been previously obtained in similar IE tasks on molecular biology.

The good results achieved by hybrid methods using machine learning and handcraft patterns show that efficient adaptation of generic methods to the task could rely not only on machine learning, but also on alternative approaches. This observation may also be true for the extraction of n-ary events from binary relations where rewriting rules may complement machine learning methods. This may be particularly appropriate for relatively small corpora as SeeDev, which belongs to a domain where a trade-off has to be found between the time needed for the training corpus annotation and the time needed for the manual development of dedicated rules for the IE method.

Acknowledgments

This work was supported by the French National Institute for Agricultural Research (Inra), the Center for Data Science (CDS), the OpenMinTeD project (EC/H2020-EINFRA 654021) and the Institute for Living Systems (IMSV), funded by the IDEX Paris-Saclay ANR-11-IDEX-0003-02. The IJPB benefits from the support of the Labex Saclay Plant Sciences- SPS (ANR-10-LABX-0040-SPS).

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC bioinformatics*, 9(11):1.

B Alberts, A Johnson, J Lewis, P Walter, M Raff, and K Roberts. 2002. Molecular biology of the cell 4th edition: International student edition.

Sophia Ananiadou, Paul Thompson, Raheel Nawaz, John McNaught, and Douglas B Kell. 2014. Event-based text mining for biology and functional genomics. *Briefings in functional genomics*, page elu015.

Mouhamadou Ba and Robert Bossy. 2016. Interoperability of corpus processing work-flow engines: the case of alvisnlp/ml in openminted. In Richard Eckart de Castilho, Sophia Ananiadou, Thomas Margoni, Wim Peters, and Stelios Piperidis, editors, *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016*, pages 15–18, Portoroz, Slovenia, May. European Language Resources Association (ELRA).

Robert Bossy, Julien Jourde, Alain-Pierre Manine, Philippe Veber, Erick Alphonse, Maarten Van De Guchte, Philippe Bessières, and Claire Nédellec. 2012. Bionlp shared task-the bacteria track. *BMC bioinformatics*, 13(11):1.

Robert Bossy, Wiktoria Golik, Zorana Ratkovic, Dialekti Valsamou, Philippe Bessières, and Claire Nédellec. 2015. Overview of the gene regulation network and the bacteria biotope tasks in bionlp'13 shared task. *BMC bioinformatics*, 16(10):1.

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*, pages 724–731. Association for Computational Linguistics.

Estelle Chaix, Bertrand Dubreucq, Dialekti Valsamou, Abdelhak Fatihi, Louise Deléger, Robert Bossy, Pierre Zweigenbaum, Philippe Bessières, Loic Lepiniec, and Claire Nédellec. 2016. Annotation guidelines bionlp-st 2016 seedev task. Technical report, INRA.

Stephen Clark and James R Curran. 2007. Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4):493–552.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics.

Martin Gerner, Goran Nenadic, and Casey M Bergman. 2010. Linnaeus: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1.

Wiktoria Golik, Robert Bossy, Zorana Ratkovic, and Claire Nédellec. 2013. Improving term extraction with linguistic analysis in the biomedical domain. In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing13), Special Issue of the journal Research in Computing Science*, pages 24–30.

Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. 2013a. The genia event extraction shared task, 2013 edition-overview. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15. Association for Computational Linguistics.

Jung-Jae Kim, Xu Han, Vivian Lee, and Dietrich Rebholz-Schuhmann. 2013b. Gro task: Populating the gene regulation ontology with events and relations. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 50–57.

DKCD Manning. 2003. Natural language parsing. *Advances in neural information processing systems*, 15:3.

Andrew McCallum, Dayne Freitag, and Fernando CN Pereira. 2000. Maximum entropy markov models for information extraction and segmentation. In *Icml*, volume 17, pages 591–598.

Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature forest models for probabilistic hpsg parsing. *Computational Linguistics*, 34(1):35–80.

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. Citeseer.

Frédéric Papazian, Robert Bossy, and Claire Nédellec. 2012. Alvisae: a collaborative web text annotation editor for knowledge acquisition. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 149–152. Association for Computational Linguistics.

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. *BMC bioinformatics*, 16(10):1.

Monica Santos-Mendoza, Bertrand Dubreucq, Sébastien Baud, François Parcy, Michel Caboche, and Loïc Lepiniec. 2008. Deciphering gene regulatory networks that control seed development and maturation in arabidopsis. *The Plant Journal*, 54(4):608–620.

Dóra Szakonyi, Sofie Van Landeghem, Katja Baerenfaller, Lieven Baeyens, Jonas Blomme, Rubén Casanova-Sáez, Stefanie De Bodt, David Esteve-Bruna, Fabio Fiorani, Nathalie Gonzalez, et al. 2015. The knownleaf literature curation system captures knowledge about arabidopsis leaf growth and development and facilitates integrated data mining. *Current Plant Biology*, 2:1–11.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Junichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on Informatics*, pages 382–392. Springer.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2012. Sr4gn: a species recognition software tool for gene normalization. *PloS one*, 7(6):e38460.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.