Ontology-based Extraction of Structured Information from Publications on Preclinical Experiments for Spinal Cord Injury Treatments

Benjamin Paassen*, Andreas Stöckel*, Raphael Dickfelder*, Jan Philip Göpfert*, Tarek Kirchhoffer§, Nicole Brazda‡,§, Hans Werner Müller‡,§, Roman Klinger*, Matthias Hartung*, Philipp Cimiano*[1]

*Semantic Computing Group, CIT-EC, Bielefeld University, 33615 Bielefeld, Germany
‡Molecular Neurobiology, Neurology, HHU Düsseldorf, 40225 Düsseldorf, Germany
§Center for Neuronal Regeneration, Life Science Center, 40225 Düsseldorf, Germany

{bpaassen,astoecke,rdickfel,jgoepfert}@techfak.uni-bielefeld.de
tarek.kirchhoffer@cnr.de, {nicole.brazda,hanswerner.mueller}@uni-duesseldorf.de
{rklinger,mhartung,cimiano}@cit-ec.uni-bielefeld.de

Abstract

Preclinical research in the field of central nervous system trauma advances at a fast pace, currently yielding over 8,000 new publications per year, at an exponentially growing rate. This amount of published information by far exceeds the capacity of individual scientists to read and understand the relevant literature. So far, no clinical trial has led to therapeutic approaches which achieve functional recovery in human patients.

In this paper, we describe a first prototype of an ontology-based information extraction system that automatically extracts relevant preclinical knowledge about spinal cord injury treatments from natural language text by recognizing participating entity classes and linking them to each other. The evaluation on an independent test corpus of manually annotated full text articles shows a macro-average F_1 measure of 0.74 with precision 0.68 and recall 0.81 on the task of identifying entities participating in relations.

1 Introduction

Injury to the central nervous system of adult mammals typically results in lasting deficits, like permanent motor and sensor impairments, due to a lack of profound neural regeneration. Specifically, patients who have sustained spinal cord injuries (SCI) usually remain partially paralyzed for the rest of their lives. Preclinical research in the field of central nervous system trauma advances at fast pace, currently yielding over 8,000 new publications per year, at an exponentially growing rate, with a total amount of approximately 160,000 PubMed-listed papers today.[2]

However, translational neuroscience faces a strong disproportion between the immense preclinical research effort and the lack of successful clinical trials in SCI therapy: So far, no therapeutic approach has led to functional recovery in human patients (Filli and Schwab, 2012). As the vast amount of published information by far exceeds the capacity of individual scientists to read and understand the relevant knowledge (Lok, 2010), the selection of promising therapeutic interventions for clinical trials is notoriously based on incomplete information (Prinz et al., 2011; Steward et al., 2012).

Thus, automatic information extraction methods are needed to gather structured, actionable knowledge from large amounts of unstructured text that describe outcomes of preclinical experiments in the SCI domain. Being stored in a database, such knowledge provides a highly valuable resource enabling curators and researchers to objectively assess the prospective success of experimental therapies in humans, and supports the cost-effective execution of meta studies based on all previously published data. First steps towards such a database have already been undertaken by manually extracting the desired information from a limited number of papers (Brazda et al., 2013), which is not feasible on a large scale, though.

In this paper, we present a first prototype of an automated ontology-based information extraction system for the acquisition of structured knowledge about experimental SCI therapies. As main contributions, we point out the highly relational problem structure by describing the entity classes and relations relevant for

[1] The first four authors contributed equally.

[2] As in this query to the database PubMed (link to `http://www.ncbi.nlm.nih.gov/pubmed`), as of April 2014.

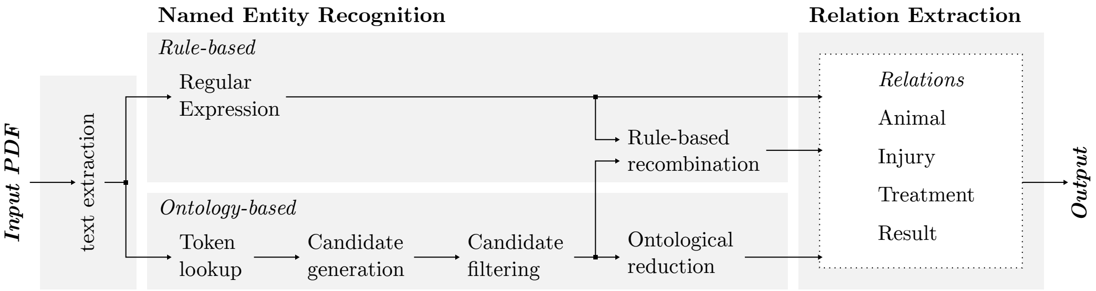

Figure 1: Workflow of our implementation, from the input PDF document to the generation of the output relations. Named entity recognition is described in Section 3.1, relation extraction in Section 3.2.

knowledge representation in the domain, and provide a cascaded workflow that is capable of extracting these relational structures from unstructured text with an average F_1 measure of 0.74.

2 Related Work

Our workflow for acquiring structured information in the domain of spinal cord injury treatments is an example of ontology-based information extraction systems (Wimalasuriya and Dou, 2010): Large amounts of unstructured natural language text are processed through a mechanism guided by an ontology, in order to extract predefined types of information. Our long-term goal is to represent all relevant information on SCI treatments in structured form, similar to other automatically populated databases in the biomedical domain, such as STRING-DB for protein-protein interactions (Franceschini et al., 2013), among others.

A strong focus in biomedical information extraction has long been on named entity recognition, for which machine-learning solutions such as conditional random fields (Lafferty et al., 2001) or dictionary-based systems (Schuemie et al., 2007; Hanisch et al., 2005; Hakenberg et al., 2011) are available which tackle the respective problem with decent performance and for specific entity classes such as organisms (Pafilis et al., 2013) or symptoms (Savova et al., 2010; Jimeno et al., 2008). A detailed overview on named entity recognition, covering other domains as well, can be found in Nadeau and Sekine (2007).

The use case described in this paper, however, involves a highly relational problem structure in the sense that individual facts or relations have to be aggregated in order to yield accurate, holistic domain knowledge, which corresponds most closely to the problem structure encountered in event extraction, as triggered by the ACE program (Doddington et al., 2004; Ji and Grishman, 2008; Strassel et al., 2008), and the BioNLP shared task series (Nedellec et al., 2013; Tsujii et al., 2011; Tsujii, 2009). General semantic search engines in the biomedical domain mainly focus on isolated entities. Relations are typically only taken into account by co-occurrence on abstract or sentence level. Examples for such search engines include GoPubMed (Doms and Schroeder, 2005), SCAIView (Hofmann-Apitius et al., 2008), and GeneView (Thomas et al., 2012).

With respect to the extraction methodology, our work is similar to Saggion et al. (2007) and Buitelaar et al. (2008), in that a combination of gazetteers and extraction rules is derived from the underlying ontology, in order to adapt the workflow to the domain of interest. A schema in terms of a reporting standard has recently been proposed by the MIASCI-consortium (Lemmon et al., 2014, Minimum Information About a Spinal Cord Injury Experiment). To the best of our knowledge, our work is the first attempt at automated information extraction in the SCI domain.

3 Method and Architecture

An illustration of the proposed workflow is shown in Figure 1. Based on the unstructured information management architecture (UIMA, Ferrucci and Lally (2004)), full text PDF documents serve as input to the workflow. Plain text and structural information are extracted from these documents using Apache PDFBox[3].

The proposed system extracts *relations* which we define as templates that contain slots, each of which is to be filled by an instance of a particular entity class (*cf.* Table 1). At the same time, a particular instance can be a filler for different slots (*cf.* Figure 2). We argue that a relational approach is essential to information extraction in the SCI domain as (i) many instances of entity classes found in the text do not convey relevant

[3] Apache PDFBox – A Java PDF Library `http://pdfbox.apache.org/`

Relation	Entity Class	Example	Method	Resource	Count
	Integer	"42", "2k", "1,000"	R	Regular Expressions	
	Float	"4.23", "$8.12 \cdot 10^{-8}$"	R	Regular Expressions	
	Roman Number	"XII", "MCLXII"	R	Regular Expressions	
	Word Number	"seventy-six"	O	*Word Number List*	99
	Range	"2-4"	R	QTY + PARTICLE + QTY	
	Language Quantifier	"many", "all"	O	*Quantifier List*	11
	Time	"2 h", "14 weeks"	R	QTY + TIME UNIT	
	Duration	"for 2h"	R	PARTICLE + TIME	
Animal	**Organism**	"dog", "rat", "mice"	O	NCBI Taxonomy	67657
	Laboratory Animal	"Long-Evans rats"	O	*Special Laboratory Animals*	5
	Sex	"male", "female"	O	*Gender List*	2
	Exact Age	"14 weeks old"	R	TIME + AGE PARTICLE	
	Age	"adult", "juvenile"	O	*Age Expressions*	2
	Weight	"200 g"	R	QTY + WEIGHT UNIT	
	Number	"44", "seventy-six"	R	QTY	
Injury	**Injury Type**	"compression"	O	*Injury Type List*	7
	Injury Device	"NYU Impactor"	O	*Injury Device List*	21
	Vertebral Position	"T4", "T8-9"	R	Regular Expressions	
	Injury Height	"cervical", "thoracic"	O	*Injury Height Expressions*	4
Treatment	**Drug**	"EPO", "inosine"	O	MeSH	14000
	Delivery	"subcutaneous", "i.v."	O	*Delivery Dictionary*	34
	Dosage	"14 ml/kg"	R	QTY + UNIT	
Result	**Investigation Method**	"walking analysis"	O	*Method List*	117
	Significance	"significant"	O	*Significance Quantifiers*	2
	Trend	"decreased", "improved"	O	*Trend Dictionary*	4
	p Value	"p < 0.05"	R	P + QTY	4

Table 1: A detailed list of relations and the entity classes whose instances are valid slot fillers for them. Examples for instances of each entity class are also shown, as well as the extraction method, and resources used for extraction. Instances are either extracted from the text using regular expressions (R) or on a lookup in our ontology database (O). Resources in *italics* were specifically created for this application, resources in SMALL CAPITALS are regular expression-based recombinations of other entities. Entity classes in bold face are *required* arguments for relation extraction (*cf.* Section 3.2). The count specifies the number of elements in the respective resource.

information on their own, but only in combination with other instances (*e. g.*, surgical devices mentioned in the text are only relevant if used to inflict a spincal cord injury to the animals in an experimental group), and (ii) a holistic picture of a preclinical experiment can only be captured by aggregating several relations (*e. g.*, a certain p value being mentioned in the text implies a particular treatment of one group of animals to be significantly different from another treatment of a control group).

We take four relations (*Animal*, *Injury*, *Treatment* and *Result*) into account which capture the semantic essence of a preclinical experiment: Laboratory animals are injured, then treated and the effect of the treatment is measured. Table 1 provides an overview of all entity classes and relations. The workflow consists of two steps: Firstly, rule- and ontology-based named entity recognition (NER) is performed (*cf.* Section 3.1). Secondly, the pool of entities recognized during NER serves as a basis for relation extraction (*cf.* Section 3.2).

3.1 Ontology-based Named Entity Recognition

We store ontological information in a relational database as a set of directed graphs, accompanied by a dictionary for efficient token lookup. Each entity is stored with possible linguistic surface forms (*e. g.*, "Wistar rats" as a surface form of the *Wistar rat* entity from the class *Laboratory Animal*). Each surface form s is tokenized (on white space and non-alphanumeric symbols, including transformation to lowercase, *e. g.*, leading to tokens "wistar" and "rats") and normalized (stemming, removal of special characters and stop words) resulting in a set of *dictionary keys* (*e. g.*, "wistar" and "rat"). The resources used as content for the ontology are shown in Table 1. We use specifically crafted resources for our use case[4] as well as the

[4] Resources built specifically are made publicly available at `http://opensource.cit-ec.de/projects/scie`

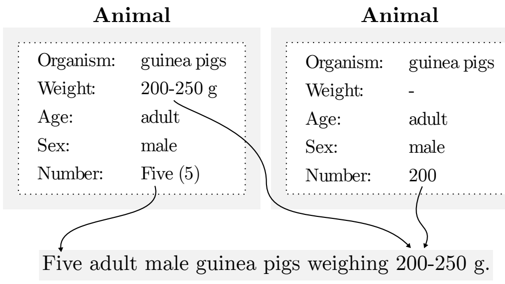

Figure 2: Two example instances of the *Animal* relation that can be generated from the same text. Given its entity class, the number 200 is a valid filler for the 'number' slot as well as the 'weight' slot. Both candidates are generated and ranked according to their probability (*cf.* Equation 4). The manually defined constraints of p_{sem} ensure that 200 cannot fill both slots at the same time.

NCBI taxonomy[5] and the Medical Subject Headings[6] (MeSH). The process of ontology-based NER consists of (i) *token lookup* in the dictionary, (ii) *candidate generation*, (iii) *probabilistic candidate filtering* and (iv) *ontological reduction* (*cf.* Figure 1).

Token lookup. For each token t in the document, the corresponding surface form tokens $\mathbf{s}_t$ are retrieved from the database. A *confidence value* p_{conf} based on the Damerau-Levenshtein-Distance without swaps (dld, Damerau (1964)) is calculated as

$$p_{\text{conf}}(t, \mathbf{s}_t) := \max\left\{0, 1 - \min_{t' \in \mathbf{s}_t} \frac{\text{dld}(t', t)}{|t'|}\right\}, \tag{1}$$

where $|t|$ denotes the number of characters in token t. Assuming to find $t =$ "rat" in the text with the according surface form $\mathbf{s}_t =$ ("wistar", "rats"), $p_{\text{conf}}(t, \mathbf{s}_t) = 1 - \frac{1}{4} = 0.75$. Tokens with $p_{\text{conf}} < 0.5$ are discarded.

Candidate generation. A candidate $\mathbf{h}$ for matching the surface form tokens $\mathbf{s}_\mathbf{h}$ is a list of tokens $(t_1^\mathbf{h}, \ldots, t_n^\mathbf{h})$ from the text. Candidates are constructed using all possible combinations of matching tokens for each surface form token (as retrieved above). To keep this tractable, we restrict the search space to combinations with the proximity $d(t_k^h, t_\ell^h) \leq 9$ for all $t_k^h, t_\ell^h \in \mathbf{h}$, where $d(u, v) := N_W(u, v) + 3 \cdot N_S(u, v) + 10 \cdot N_P(u, v)$ models the distance between two tokens u and v in the text with N_W, N_S, N_P denoting the number of words, sentences and paragraphs between u and v. In our example, a candidate would be $\mathbf{h} =$ ("rat").

Candidate filtering. For a candidate $\mathbf{h}$ and the surface form tokens $\mathbf{s}_\mathbf{h}$ it refers to, we calculate a total *match probability*, taking into account the distance $d(u, v)$ of all tokens in the candidate, the confidence $p_{\text{conf}}(t', \mathbf{s}_\mathbf{h})$ that the token actually belongs to the surface form, and the ratio $\sum_{t' \in \mathbf{h}} |t'| / \sum_{t \in \mathbf{s}_\mathbf{h}} |t|$ of the surface form tokens covered by the candidate:

$$p_{\text{match}}(\mathbf{h}, \mathbf{s}_\mathbf{h}) = \frac{1}{\sum_{t \in \mathbf{s}_\mathbf{h}} |t|} \max_{t \in \mathbf{h}} \sum_{t' \in \mathbf{h}} \left(p_{\text{dist}}^3(t, t') \cdot p_{\text{conf}}(t', \mathbf{s}_\mathbf{h}) \cdot |t'|\right), \tag{2}$$

$$\text{where } p_{\text{dist}}^\sigma(u, v) := \exp\left(-\frac{d(u, v)^2}{2\sigma^2}\right) \tag{3}$$

models the confidence that two tokens u and v belong together given their distance in the text. In our example of the candidate $\mathbf{h} =$ ("rat") with the surface form tokens $\mathbf{s}_\mathbf{h} =$ ("wistar", "rats") is $p_{\text{match}}(\mathbf{h}, \mathbf{s}_\mathbf{h}) = 1 \cdot 0.75 \cdot \frac{3}{6+4} = 0.225$. Candidates with $p_{\text{match}} < 0.7$ are discarded. The resulting set of all recognized candidates is denoted with H.

Ontological reduction. As the algorithm ignores the hierarchical information provided by the ontologies, we may obtain overlapping matches for ontologically related entities. Therefore, in case of overlapping entities that are related in an "is a" relationship in the ontology, only the more specific one is kept. Assume for instance the candidates "Rattus norvegicus" and "Rattus norvegicus albus", where the latter is more specific and therefore accepted.

3.2 Relation Extraction

We frame *relation extraction* as a template filling task such that each slot provided by a relation has to be assigned a filler of the correct entity class. Entity classes for the four relations of interest are shown in

[5] Sayers et al. (2012), database limited to vertebrates: `http://www.ncbi.nlm.nih.gov/taxonomy/?term=txid7742[ORGN`

[6] Lipscomb (2000), excerpt of drugs from Descriptor and Supplemental: `https://www.nlm.nih.gov/mesh/`

Table 1, where *required* slots are in bold face, whereas all other slots are *optional*.

The slot filling process is based on testing all combinations of appropriate entities taking into account their proximity and additional constraints. In more detail, we define the set of all recognized relations $\mathcal{R}_\theta$ of a type θ as

$$\mathcal{R}_\theta = \left\{ r^\theta \in \mathcal{P}(H) \;\middle|\; \frac{p_{\text{sem}}(r^\theta)}{n^\theta} \cdot \sum_{\mathbf{h} \in r^\theta, \mathbf{h} \neq g(r^\theta)} p_{\text{match}}(\mathbf{h}, \mathbf{s}^{\mathbf{h}}) \min_{t \in \mathbf{h}, t' \in g(r^\theta)} p_{\text{dist}}^{\sigma_\theta}(t, t') > 0.2 \right\} \tag{4}$$

where $\mathcal{P}(H)$ denotes the power set over all candidates H recognized by NER. $g(r^\theta)$ returns the filler for the *required* slot of r^θ, p_{match} and p_{dist} are defined as in Section 3.1 and p_{sem} implements manually defined constraints on r^θ: A wrongly typed filler h for one slot of r^θ leads to $p_{\text{sem}}(r^\theta) = 0$, as does a negative number in the *Number* slot of the *Animal* relation. Animal Numbers larger than 100 or Animal Weights smaller than 1 g or larger than 1 t are punished. All other cases lead to $p_{\text{sem}}(r^\theta) = 1$. Note that $p_{\text{match}}(\mathbf{h}, \mathbf{s}^{\mathbf{h}}) = 1$ for candidates h retrieved by rule-based entity recognition. Further, we set $\sigma_{\text{Animal}} = \sigma_{\text{Treatment}} = 6$, $\sigma_{\text{Injury}} = 10$ and $\sigma_{\text{Result}} = 15$.

4 Experiments

4.1 Data Set

The workflow is evaluated against an independent, manually annotated corpus of 32 complete papers which contain 1186 separate annotations of entities, produced by domain experts[7]. Information about relations is not provided in the corpus. Only entities which participate in the description of the preclinical experiment are marked. The frequencies of annotations among the different classes are shown in Table 2.

4.2 Experimental Settings

We evaluate the system with regard to two different tasks: *extraction* ("Is the approach able to extract relevant information from the text, without regard to the exact location of the information?") and *annotation* ("Is the system able to annotate relevant information at the correct location as indicated by medical experts?"). Furthermore, we distinguish between an *all instances* setting, where we consider all instances independently, and a *fillers only* setting, where only those annotations in the system output are considered, that are fillers in a relation (*i.e.* the fillers only-setting evaluates a subset of the all instances-setting). The relation extraction procedure is not evaluated separately. For each setting, we report precision, recall, and F_1 measure.

Overall	1186
Organism	58
Weight	32
Sex	33
Age	17
Injury Height	35
Injury Type	62
Injury Device	23
Drug	134
Dosage	106
Delivery	70
Investigation Method	129
Trend	219
Significance	137
p Value	131

Table 2: The number of annotations in our evaluation set for each entity class.

Taking the architecture into account, we have the following hypotheses: (i) For the *all instances* setting we expect high recall, but low precision. (ii) For the *fillers only* setting, precision should increase notably. (iii) Comparing the *all entities* and the *fillers only* setting, recall should remain at the same level. We therefore expect the *extraction* task to be simpler than the *annotation* task: For any information to be annotated at the correct position, it must have been extracted correctly. On the other hand, information that has been extracted correctly, can still be found at a 'wrong' location in the text. Thus, we expect a drop of precision and recall when moving from *extraction* to *annotation*.

4.3 Results

The results are presented in Table 3: For each relation mentioned in Section 3, and the entity classes participating in it, we report precision, recall and F_1-measure[8]. This is done for all four combinations of setting and task. For each relation we also provide the macro-average of precision, recall and F_1-measure over all entity classes considered in that relation and the overall average.

[7] Performed in Protégé http://protege.stanford.edu/ with the plug-in Knowtator http://knowtator.sourceforge.net/ (Ogren, 2006)

[8] Note that *VertebralPosition* and *InjuryHeight* are merged in the result table, as are *Organism* and *Laboratory Animal* and *Age* and *Exact Age*. The *Animal Number* was excluded from the evaluation as it has not been annotated in our evaluation set.

Task	Extraction						Annotation					
Setting	All Instances			Fillers Only			All Instances			Fillers Only		
Entity Class	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Overall Average	0.58	0.95	0.72	0.68	0.81	0.74	0.13	0.77	0.22	0.21	0.51	0.30
***Animal* Average**	0.62	0.99	0.76	0.82	0.94	0.87	0.12	0.91	0.21	0.31	0.81	0.44
Organism	0.41	1.00	0.58	0.88	0.90	0.89	0.02	1.00	0.04	0.24	0.66	0.35
Weight	0.20	1.00	0.33	0.52	0.94	0.67	0.08	0.97	0.15	0.49	0.91	0.64
Sex	0.85	0.99	0.91	0.87	0.98	0.92	0.18	0.94	0.30	0.26	0.94	0.41
Age	1.00	0.95	0.97	1.00	0.93	0.96	0.19	0.71	0.30	0.23	0.71	0.35
***Injury* Average**	0.63	0.94	0.76	0.74	0.75	0.75	0.12	0.72	0.21	0.18	0.38	0.24
Injury Height	0.42	0.98	0.59	0.56	0.74	0.64	0.10	0.91	0.18	0.24	0.51	0.33
Injury Type	0.70	0.91	0.79	0.81	0.73	0.77	0.07	0.48	0.12	0.18	0.35	0.24
Injury Device	0.78	0.93	0.85	0.86	0.79	0.82	0.20	0.77	0.32	0.11	0.28	0.16
***Treatment* Average**	0.45	0.91	0.61	0.53	0.78	0.63	0.14	0.72	0.23	0.19	0.54	0.28
Drug	0.10	0.98	0.18	0.24	0.69	0.36	0.01	0.74	0.02	0.10	0.42	0.16
Dosage	1.00	0.81	0.90	1.00	0.76	0.86	0.30	0.52	0.38	0.32	0.46	0.38
Delivery	0.26	0.95	0.41	0.34	0.89	0.49	0.11	0.89	0.20	0.15	0.74	0.25
***Result* Average**	0.59	0.93	0.72	0.60	0.75	0.67	0.13	0.71	0.22	0.15	0.30	0.20
Investigation Method	0.29	0.96	0.45	0.27	0.79	0.40	0.03	0.66	0.06	0.02	0.16	0.04
Trend	0.37	0.91	0.53	0.44	0.78	0.56	0.06	0.63	0.11	0.07	0.27	0.11
Significance	0.70	0.90	0.79	0.70	0.71	0.70	0.17	0.69	0.27	0.22	0.39	0.28
p Value	1.00	0.96	0.98	1.00	0.71	0.83	0.27	0.86	0.41	0.30	0.36	0.33

Table 3: The macro-averaged evaluation results for each class given in precision, recall and F_1 measure.

For the *extraction* task with *all instances* setting, recall is close to 100% for all entity classes considered in the *Animal* relation. It is 81% for Dosages. The rule-based recognition for Dosages (as for Ages and p Values) is very precise: All recognized entities have been annotated by medical experts somewhere in the document. This strong difference between entity classes can be observed in the *annotation* task and the *fillers only* setting as well: The best average performance in F_1-measure is achieved for entity classes that are part of the Animal relation. Precision is best for Dosages, Ages and p Values.

The recall for the *all instances* setting is high in both the extraction and in the annotation task. However, the number of annotated instances (29,628 annotations in total) is about 25 times higher than the number of expert annotations, which leads to low precision especially in the annotation task. For the *fillers only* setting, the number of annotations decreases dramatically (to 4069 annotations); at the same time, precision improves. Regarding the comparison of both tasks, precision and recall are both notably lower in the annotation task, for the *all entities* setting, as well as for the *fillers only* setting. The overall recall is lower by 14 percentage points (pp) in the extraction task and by 26 pp in the annotation task when considering the *fillers only* setting. The decrease is most pronounced for Investigation Methods in the annotation task with a drop of 50 pp.

4.4 Discussion

The results are promising for named entity recognition. Recall is close-to-perfect in the *extraction* task and acceptable in the *annotation* task. The results for relation extraction leave space for improvement: An increase in precision can be observed but the decrease in recall is too substantial. The *Animal* relation is an exception, where an increase in F_1 measure is observed for the *fillers only* setting for nearly all entity classes, leading to 0.87 F_1 for *Animals* in the *extraction* task.

An error analysis revealed that for the *fillers only* setting, most false positives (55%) are due to the fact that the medical experts did not annotate *all* occurrences of the correct entity, but only one or a few. 18% are due to ambiguities of surface forms (for instance the abbreviation “it” for “intrathecal” leads to many false positives). Regarding false negatives, 41% are due to missing entries in our ontology database and further 26% are caused by wrong treatment of characters (mostly wrong transcriptions of characters from the PDF).

5 Conclusion and Outlook

We described the challenge of extracting relational descriptions about preclinical experiments on spinal cord injury from scientific literature. To tackle that challenge, we introduced a cascaded approach of named entity recognition, followed by relation extraction. Our results show that the first step can be achieved by relying strongly on domain-specific ontologies. We show that modeling relations as aggregated entities, and extracting them using a distance filtering principle combined with domain specific knowledge, yields promising results, specifically for the *Animal* relation.

Future work will focus on improving the recognition at the correct position in the text. This is a prerequisite to actually tackle and evaluate the relation extraction not only on the basis of detected participating entities. Therefore, improved relation detection approaches will be implemented which relax the assumption that relevant entities are found close-by in the text. In addition, we will relax the assumption that different slots of the annotation are all equally important. Finally, we will address aggregation beyond individual relations in order to allow for a fully accurate holistic assessment of experimental therapies.

Our system offers a semantic analysis of scientific papers on spinal cord injuries. This lays groundwork for populating a comprehensive semantic database on preclinical studies of SCI treatment approaches as described by Brazda et al. (2013), laying ground and supporting transfer from preclinical to clinical knowledge in the future.

References

N. Brazda, M. Kruse, F. Kruse, T. Kirchhoffer, R. Klinger, and H.-W. Müller. 2013. The CNR preclinical database for knowledge management in spinal cord injury research. *Abstracts of the Society of Neurosciences*, 148(22).

P. Buitelaar, P. Cimiano, A. Frank, M. Hartung, and S. Racioppa. 2008. Ontology-based information extraction and integration from heterogeneous data sources. *Int. J. Hum.-Comput. Stud.*, 66(11):759–788.

F. J. Damerau. 1964. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM*, 7(3):171–176, March.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. 2004. The Automatic Content Extraction (ACE) program: tasks, data, and evaluation. In *Proceedings of LREC 2004*, pages 837–840.

A. Doms and M. Schroeder. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*, 33(Web Server issue):W783–W786, Jul.

D. Ferrucci and A. Lally. 2004. Building an example application with the Unstructured Information Management Architecture. *IBM Systems Journal*, 43(3):455–475.

L. Filli and M. E. Schwab. 2012. The rocky road to translation in spinal cord repair. *Ann Neurol*, 72(4):491–501.

A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, 41(Database issue):D808–D815, Jan.

J. Hakenberg, M. Gerner, M. Haeussler, I. Solt, C. Plake, M. Schroeder, G. Gonzalez, G. Nenadic, and C. M. Bergman. 2011. The GNAT library for local and remote gene mention normalization. *Bioinformatics*, 27(19):2769–2771, Oct.

D. Hanisch, K. Fundel, H.-T. Mevissen, R. Zimmer, and J. Fluck. 2005. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6 Suppl 1:S14.

M. Hofmann-Apitius, J. Fluck, L. Furlong, O. Fornes, C. Kolarik, S. Hanser, M. Boeker, S. Schulz, F. Sanz, R. Klinger, T. Mevissen, T. Gattermayer, B. Oliva, and C. M. Friedrich. 2008. Knowledge environments representing molecular entities for the virtual physiological human. *Philos Trans A Math Phys Eng Sci*, 366(1878):3091–3110, Sep.

H. Ji and R. Grishman. 2008. Refining Event Extraction through Cross-Document Inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio, June. Association for Computational Linguistics.

A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9 Suppl 3:S3.

J. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML 2001*, pages 282–289. Morgan Kaufmann.

V. P. Lemmon, A. R. Ferguson, P. G. Popovich, X.-M. Xu, D. M. Snow, M. Igarashi, C. E. Beattie, J. L. Bixby et al. 2014. Minimum Information About a Spinal Cord Injury Experiment (MIASCI) – a proposed reporting standard for spinal cord injury experiments. *Neurotrauma*. in press.

C. E. Lipscomb. 2000. Medical Subject Headings (MeSH). *Bull Med Libr Assoc*, 88(3):265–266, Jul.

C. Lok. 2010. Literature mining: Speed reading. *Nature*, 463(7280):416–418, Jan.

D. Nadeau and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

C. Nedellec, R. Bossy, J.-D. Kim, J. jae Kim, T. Ohta, S. Pyysalo, and P. Zweigenbaum, editors. 2013. *Proceedings of the BioNLP Shared Task 2013 Workshop*. Association for Computational Linguistics, Sofia, Bulgaria, August.

P. V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings NAACL/HLT 2006*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

E. Pafilis, S. P. Frankild, L. Fanini, S. Faulwetter, C. Pavloudi, A. Vasileiadou, C. Arvanitidis, and L. J. Jensen. 2013. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS One*, 8(6):e65390.

F. Prinz, T. Schlange, and K. Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*, 10(9):712, Sep.

H. Saggion, A. Funk, D. Maynard, and K. Bontcheva. 2007. Ontology-Based Information Extraction for Business Intelligence. In K. A. et al., editor, *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 843–856. Springer.

G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–513.

E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. 2012. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 40(Database issue):D13–D25, Jan.

M. Schuemie, R. Jelier, and J. Kors. 2007. Peregrine: lightweight gene name normalization by dictionary lookup. In *Proceedings of the Biocreative 2 workshop 2007*, page 131–140, Madrid, Spain, April.

O. Steward, P. G. Popovich, W. D. Dietrich, and N. Kleitman. 2012. Replication and reproducibility in spinal cord injury research. *Exp Neurol*, 233(2):597–605, Feb.

S. Strassel, M. Przybocki, K. Peterson, Z. Song, and K. Maeda. 2008. Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2706–2709.

P. Thomas, J. Starlinger, A. Vowinkel, S. Arzt, and U. Leser. 2012. GeneView: a comprehensive semantic search engine for PubMed. *Nucleic Acids Res*, 40:W585–W591, Jul.

J. Tsujii, J.-D. Kim, and S. Pyysalo, editors. 2011. *Proceedings of BioNLP Shared Task 2011 Workshop*. Association for Computational Linguistics, Portland, Oregon, USA, June.

J. Tsujii, editor. 2009. *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*. Association for Computational Linguistics, Boulder, Colorado, June.

D. C. Wimalasuriya and D. Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323.