MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-521

Scholarly Information Extraction Is Going to Make a Quantum Leap with PubMed Central (PMC)® —But Moving from Abstracts to Full Texts Seems Harder than Expected

Franz Matthies and Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität Jena, Jena 07743, Germany

Abstract

With the increasing availability of complete full texts (journal articles), rather than their surrogates (titles, abstracts), as resources for text analytics, entirely new opportunities arise for information extraction and text mining from scholarly publications. Yet, we gathered evidence that a range of problems are encountered for full-text processing when biomedical text analytics simply reuse existing NLP pipelines which were developed on the basis of abstracts (rather than full texts). We conducted experiments with four different relation extraction engines all of which were top performers in previous BioNLP Event Extraction Challenges. We found that abstract-trained engines loose up to 6.6% F-score points when run on full-text data. Hence, the reuse of existing abstract-based NLP software in a full-text scenario is considered harmful because of heavy performance losses. Given the current lack of annotated full-text resources to train on, our study quantifies the price paid for this short cut.

Keywords:

Natural Language Processing; Information Storage and Retrieval; Information Extraction

Introduction

While abstracts contain only the major results of the corresponding journal article in a highly condensed manner, the full-text body of scholarly publications makes accessible all single pieces of information from scientific studies. Since the demand for this maximum level of information is also high in the life sciences, the NLM launched a full-text collection for a subset of PUBMED abstracts, PUBMED CENTRAL (PMC),[1] which currently (April 2017) archives 4.3 million articles and is growing at high speed. Given such a bulky resource, unlimited access to the information contained in scholarly full texts seems both a realistic and rewarding goal, since a great deal of relevant biomedical information is only contained in the full-text portions of scientific articles and is not mentioned at all in the corresponding abstracts of the full texts [1–3] – so the information gain from processing full texts can be enormous.

Tanabe and Wilbur [4] were the first to hint at technical problems of dealing with special non-ASCII characters, tables and figures embedded in full texts. Going beyond low-level technicalities, within the BIOCREATIVE Gene Normalization Challenge, for the first time, evidence was gathered that performance dropped significantly when tested on full texts [5] instead of abstracts [6].

Cohen *et al.* [7] substantiated these warnings that the processing of full texts will be more than challenging by investigating their intrinsic properties. They conducted an empirical study in which they investigated different structural and linguistic properties of abstracts and their corresponding full texts. They found longer sentences in the full texts than in the abstracts and much heavier use of parenthesized material (e.g., abbreviations, citations, data values, figure/table pointers, etc.) in the full texts than in the abstracts. Both phenomena make full texts much harder to parse than abstracts. Syntax-wise they also gathered evidence that the incidence of conjunctions, passives, pronominal anaphora, as well as sentence complexity/readability were markedly different from full texts to abstracts. However, syntactic parsing using the Stanford Lexicalized Parser yielded no statistically relevant difference between both genres (taking ParsEval's metrics for bracket recall and tag accuracy). Yet, POS tagging was more accurate in abstracts than in full texts. Analysis errors caused by the increased syntactic complexity of full text articles (compared with abstracts) were also recognized by Tudor *et al.* [8]. McIntosh and Curran [9] further point out that coreference relations (anaphora) between sentences play an important role in full texts. Semantics-wise, Cohen *et al.* found that the distribution of named entities such as genes/proteins, mutations, drugs, etc. differed between the two text genres as well.

Most important for our own work, Cohen *et al.* also investigated the impact these differences had on the performance of information extraction tools. They found that three common gene mention recognition systems (BANNER, ABNER and LINGPIPE) performed much worse on full texts than they did on abstracts – F-measures were generally about 10 points higher on the abstracts than on the full text portions. Consequently, the authors advocated retraining models on full texts.

If this suggestion is taken seriously, annotated full-text corpora have to be supplied. Given the large volume of utterances contained in full texts, this is an enormously resource-consumptive task. Cohen *et al.*'s study is based on their self-developed CRAFT full text corpus [10] which reflects the findings for full text phenomena from [7]. It consists of 97 (67 publically released) full texts, plus corresponding abstracts (comprising nearly 31k sentences or 800k text tokens), dealing with mouse genomics. This corpus was annotated for part-of-speech (POS) and syntactic parsing data, as well as proteins/genes as named entities [11].

[1] https://www.ncbi.nlm.nih.gov/pmc/

Table 1 - Overview of relation extraction systems used for our experiments and their ranking in the BioNLP Shared Task 2013 on the GENIA Task (in parentheses, results from the BioNLP Shared Task 2009, if applicable)

Name	Type	Ranking		Group	Reference
Tees	ML (SVM)	2.	(1.)	Turku, FI	[24], [25]
BioSem	ML + rule-based	3.	(n.a.)	Amsterdam/Rotterdam, NL	[26], [27]
HDS4NLP	ML (SVM)	6./1.	(n.a.)	Compiègne, FR	[28]
JReX	ML (SVM, MaxEnt)	n.a.	(2.)	Jena, DE	[14], [29]

The authors carefully discuss the large variation in testing conditions, e.g., the types of named entity taggers (Banner, Abner and LingPipe), the types of models being considered (trained on the abstract-based BioCreative II, NLPBA and Genia and the full-text-based Craft corpora where, in addition, the full set is distinguished from a development set), different definitions of the named entity type 'protein' and different matching criteria for gene mentions). Still, the Craft study reveals inconsistent results for the 'gene' named entity recognition problem. For the LingPipe system, gains for retraining on the full-text corpus amount to 0.18 F-score points, but Abner's performance drops after retraining (increased precision comes at the cost of substantial recall losses). Banner's results are already competitive with the out-of-the-box model so that no retraining on Craft-style full text corpora is required. The authors summarize their results that "retraining the gene mention recognition systems unfortunately did not show much improvement" [10, p. 21] but anticipate a significant improvement if the learning problem will be rephrased (see also [10, p. 21]). Note that Craft is not annotated for any relational information so that no empirical data could be collected for this task from this resource. The Craft study also measured the impact of sentence splitting, tokenization, POS tagging and syntactic parsing relative to the two text genres—abstracts and full texts. For parsing, e.g., substantial improvements of Craft-trained models over standard (non-biomedical) English models were found.

Another full-text corpus, ID, remedies the lack of biomedical relation annotations. It was supplied for the series of BioNLP Event Extraction Challenges starting in 2009, with the second round in 2011 [12]. ID contains 30 full texts (more than 5k sentences, 150k text tokens) annotated for biomolecular mechanisms of infectious diseases that involve associations between multiple types of molecular entities, disease-causing microorganisms and other organisms undergoing the diseases. With more than 13k named entities and 4,15k events the entity count is comparable to the 2009 Challenge, whereas the event count is only approximately one third of the Shared Task 2009 data. On this data, the top-performing systems scored at almost the same level as in 2009 (where abstracts were analyzed), with the winner system (Faust) peaking on 56% F-score (trained and evaluated on sections of ID full texts).

Since the biomedical NLP community has developed a battery of well-engineered analytic engines—starting from domain-adapted tokenizers, POS taggers and parsers to domain-specific named entity taggers and relation extractors—one is tempted to simply reuse these tools and composite pipelines on full texts in order to unlock the vast amount of still hidden information. The fact that almost all of them were developed and fine-tuned on abstracts as textual training data does not seem to be an issue here. Instead, we claim that moving from the abstract to the full-text level of analysis is by no means a free lunch. We rather stipulate that classifiers trained on abstracts (irrespective of whether they deal with named entities or relations between them) will drop significantly in performance when run on corresponding full texts due to the increased level of linguistic, structural and conceptual complexity in the latter. Thus, the potential cheap benefit of making full texts available, instead of (informationally much poorer) abstracts, is likely to disappear. —But at what rate? And, is this rate, when quantified, tolerable or not?

In this paper, we deal with the problem of trading off abstract *vs*. full-text processing for the life sciences domain, with focus on relation extraction. We thus study the effects of text genres on system performance by switching between abstract and full-text documents. We substantiate our claims by running four prominent relation extraction systems, which were top-performers in the most recent BioNLP 2013 Event Extraction Challenge [13]. This selection of systems should guarantee that despite the small 'sample size' of the four classifiers the results we achieve might cautiously be generalized for a much larger class of relation taggers in the biomedical domain.

Methods

In this section, we describe the experimental set-up of our work. First, we introduce the systems we used for relation extraction. Then, we describe the full-text resources for our experiments. From a methodological perspective, a general observation can be made. The best-performing BioNLP systems (cf. the top performers in challenge competitions such as BioCreAtive [15-17], BioNLP Shared Task [18,19], i2b2 [20,21], or DDI [22,23]) either exclusively rely on some (semi-)supervised form of machine learning (ML) techniques, or combine ML with rule- or dictionary-based approaches in terms of hybrid systems. The ML systems (or ML portions of hybrid systems) being used are thus highly dependent on their textual input for training, i.e., subsets of PubMed (Medline) abstracts annotated by humans, since they constitute the gold standards both for training and evaluating these classifiers. Unsupervised ML systems are rare and, if running against competitor systems, are usually outperformed by supervised approaches in the challenges.

Relation Extractors

For our experiments, we used four systems that performed exceptionally well in previous BioNLP Event Extraction Challenges. Another criterion of choice was technical in nature—(the code of) the systems should be easily accessible and processable without much effort in our computing environment. The latter is important because technical portability of systems is an indicator for the extramural reproducibility of results. In the following, we briefly summarize each system and point out main differences among them and their approaches to relation extraction. For a more in-depth description, we kindly refer to the respective papers cited in Table 1.

Tees

The Tees system [24] was successful in all three BioNLP Shared Tasks, achieving the first rank in 2009 and the first rank in half of the subtasks in 2011 and 2013, respectively. Since the team has also provided the code as an open source project[2] to the scientific community, it was clearly the first choice for our experiments.

Tees approaches the task of relation extraction in a linear fashion, by first detecting potential triggers in a text and, in

[2] https://github.com/jbjorne/TEES

the next step, defining valid edge candidates between entities: triggers or arguments. At this point, overlapping events are possible—that is, they share the same trigger node. This 'merged event graph' is unmerged in the next step. A Support Vector Machine (SVM) sequentially and independently solves all these steps with a linear kernel handling them as multi-class classification problems.

BioSem

The BioSem system[3] [26] stands out from the other three systems we consider here because of two interesting facts. Firstly, in the training phase, it learns rules for relation representations by means of a semantic and syntactic feature list and, secondly, it refrains from deep syntactic parsing, but rather builds its structured representation of the text from the output of a shallow parse (i.e., chunks) only. Among other decisions, this leads to an outstanding computational efficiency, paving the way to relation extraction on a larger scale. The authors estimate their system to be around 170–230 times faster for completely extracting an event from a sentence compared to state-of-the-art ML-based systems. (see [26] "Computational Performance"). The usage of rules also yields a superior precision (between 60-70% depending on the set-up) compared to the other systems we consider here.

HDS4NLP

In the official results for the Genia task, HDS4NLP [28] only achieved rank 6 (with 43.03 F-score). However, its developers identified a serious bug after disclosure of the test results, and, after fixing it, their system reportedly achieved an F-score of 51.15, outperforming the top ranked competitor Evex (50.97 F-score).[4] Since they trained the model for producing the results for the 2013 Shared Task on all documents from both the development and the training sets of the BioNLP 2011 and 2013 GE tasks, the numbers we present here differ from the ones they reported, as we only use 2013 GE task documents. However, we still achieve a much higher F-score (47.81) than the official figures, putting them among the top five performers of the challenge. Also, HDS4NLP outperforms the other three systems when evaluated on terms of the BioNLP 2009 Shared Task with 54.37 F-score by at least 2.3 points.

In comparison to the other systems used for our experiments, HDS4NLP tackles relation extraction by training a model that directly extracts pairwise structured events (and the event type they belong to) of the form (*trigger*, *argument*) rather than relying on a sequential technique (i.e., extracting triggers first and then looking for applicable arguments). A point the authors make why this might be beneficial is that the usual approach of detecting triggers in isolation could lead to contextual information loss. Running an SVM (as implemented through Python's scikit-learn environment)[5] on the sentence level addresses the problem of classifying candidate-argument pairs [26]—taken from the cross product of the sets of possible trigger tokens and arguments—in a one-*vs*.-all set-up. We will elaborate on this point in the Results Section.

JReX

The JReX system, developed in our lab, participated only in the BioNLP 2009 GE task [14] where it ranked on 2nd place among 24 teams. Since then it has undergone almost no updates. This stagnation is clearly reflected in the performance figures compared to the other three systems used in our experiments which are/were under active development. JReX incorporates manually curated dictionaries and ML methodologies (SVM, MaxEnt) to sort out associated event triggers and arguments on *trimmed* dependency graph structures, the latter being simplified dependency structures from which representational 'noise' has been eliminated (cf. [14, 29]).

Text Corpora

The textual resources we exploited were taken from the text repository of the 2013 Event Extraction Challenge for the full texts and the 2011 repository for the abstracts (cf. Table 2 for a quantitative breakdown). For the former to get a reasonable amount of material, we used training and development set, whereas for the latter we only used the train set for training as to keep the size approximately on a par with that of the full text collection.

Table 2 - Overview of Text Corpora; $mix = train \cup dev$

Items	Abstract		Full Text		
	train	test	train	mix	test
Documents	800	260	10	20	14
Sentences	7,449	2,447	2,438	5,165	3,204
Words	176,146	57,367	54,938	112,845	75,144
Events	8,597	3,182	2,817	6,016	3,270

This set-up results in performance figures for the systems that differ from those reported in the 2009 Shared Task, since they used both development and training sets as input for training. Differences in the results when replicating the Shared Task 2013 set-up (compared to the original scores) can be accredited to the fact that all the systems used 2011 (abstracts, as well) and 2013 resources for training.

Events

In order to give an idea of the quantitative scope of the text collections we used, in Table 3, we distinguish the Abstract and Full Text corpus, with counts of all

- *simple* unary relations which refer to all events constituted only of an *event trigger* (a sequence of tokens indicating an event mention) and an *argument*, a protein or gene,
- the binary *Binding* relation which can have two *arguments* (both a gene or a protein),
- *Regulation* relations, which – besides proteins or genes - can also have other events as *arguments*.

Results

In this section, we present the results the four systems achieved for all possible combinations of training and evaluation data, i.e., the cross-product of the abstract (AB) and full text (FT) material. Hence, we:

- trained models on the Abstract corpus and evaluated them on Abstracts (**AB on AB**),

Table 3 - Relation Type Counts (Test set) – The numbers for Full Texts do not incorporate relation counts for Protein modification and Ubiquitination.

Relation Type	Abstract	Full Text
Simple Relations	1,182	993
Binding Relations	347	333
Σ Relations	1,529	1,326
Regulation Relations	1,653	1,944
Total	3,182	3,270

[3] In consultation with the main developer their source code was adapted to be fully compatible with the UIMA framework and is available from: https://github.com/JULIELab/jcore-dependencies/tree/master/biosem-event-extractor

[4] Available from https://github.com/XiaoLiuAI/RUPEE

[5] http://scikit-learn.org

Table 4 - Results for all conditions in Recall/Precision/F-Score triples; bold values mark the best value for the respective set-up

Name	AB on AB	AB on FT	FT on FT	FT on AB
TEES	(**46.67**/56.74/51.40)	(39.47/50.64/45.01)	(**43.17**/57.30/**49.24**)	(**40.19**/51.57/45.18)
BIOSEM	(41.39/**70.17**/52.07)	(37.11/**63.05**/46.72)	(38.41/**67.65**/49.00)	(37.12/**71.36**/**48.83**)
HDS4NLP	(46.32/65.80/**54.37**)	(**43.77**/59.12/**50.30**)	(42.84/54.09/47.81)	(39.53/54.36/45.78)
JREX	(39.75/51.97/45.05)	(37.20/44.80/40.65)	(37.69/48.03/42.23)	(37.15/52.51/43.51)

Table 5 - Deltas in Recall/Precision/F-Score triples for comparison between different columns of Table 4

	AB on AB			FT on FT	
Name	FT on FT	AB on FT	FT on AB	FT on AB	AB on FT
TEES	(–3.50/ +0.56/ –2.16)	(–7.20/–6.10/–6.39)	(–6.48/–5.17/–6.22)	(–2.98/–5.72/–4.06)	(–3.70/–6.66/–4.23)
BIOSEM	(–2.98/–2.52/–3.07)	(–4.28/–7.12/–5.35)	(–4.27/+1.19/–3.24)	(–1.29/ +3.71/–0.17)	(–1.30/–4.60/–2.28)
HDS4NLP	(–3.48/–11.71/–6.56)	(–2.55/–6.68/–4.07)	(–6.79/–11.44/–8.59)	(–3.31/ +0.27/–2.03)	(+0.93/+5.03/ +2.49)
JREX	(–2.06/–3.94/–2.82)	(–2.55/–7.17/–4.40)	(-2.60/+0.54/–1.54)	(–0.54/ +4.48/ +1.28)	(–0.49/–3.23/–1.58)

- trained models on the Abstract corpus and evaluated them on Full Texts (**AB on FT**),
- trained models on the Full Text corpus and evaluated them on Abstracts (**FT on AB**),
- trained models on the Full Text corpus and evaluated them on Full Texts (**FT on FT**).

Table 4 shows the resulting figures in terms of Recall/ Precision/F-Score triples, whereas Table 5 depicts the deltas (differences) for different condition comparisons across set-ups. When simply comparing the systems’ intra-collection performance (cf. columns **AB on AB** *vs*. **FT on FT**), we already see an F-Score difference (Δ) in favor of the abstracts, where TEES’ FT performance is the most robust (only 2.16 F-score points loss compared with AB), whereas HDS4NLP turns out to be the most fragile system (6.56 F-score points loss compared with AB).

This is an early indicator that relations are generally harder to extract from full texts than from abstracts. However, due to the difference in the relation count between the training material for Abstracts and Full Texts this conclusion is still preliminary and will be treated with some reservation. We will come back to this issue in the Discussion Section.

A noteworthy performance drop, ranging from approximately 4.1 to 6.4 points in F-Score, can be observed consistently for all systems when evaluating abstract-trained models on full text compared to a scenario when these models were tested on abstracts only (cf. columns **AB on AB** *vs*. **AB on FT**). These deltas are already strong despite the fact that we could not test for statistical significance (see Section Discussion).

In combination, both result sets already indicate that *all* systems fail to capture relation structure encodings that are present in the full text but not in the abstract. This view might be further encouraged by looking at the performance of Full Text models evaluated on Abstract and Full Text, respectively (cf. columns **FT on FT** vs. **FT on AB**). Not only do these results show no such clear picture (the deltas are much smaller), but also does the JREX system goes the other way round (F: +1.3).

Discussion

There are some caveats that need to be made explicitly for the experimental set-up we have defined:

- **Text Genre Mix-up.** The collection we here referred to as “Full Text” does include the abstracts. We could not change this mix-up because we, obviously, do not have any access to the test set. We could have done this separation for the training phase, but this would have led to the elimination of roughly about 400 relations.
- **Training Set Imbalance.** The textual material used for training the Full Text models is lower in size than that for training the Abstract models (approximately 6,000 *vs*. 8,600 events, respectively). This could be seen as a problem for comparability of the results. Yet, we still have the strong figures where Abstract models were tested on both AB and FT, and we did further experiments on both JREX and BIOSEM where these systems were subsequently tested with more and more training material for both AB and FT on both AB and FT (data not shown). In order to get an approximate increase of 2.5 F-Score for AB training material (regardless of the testing material; AB or FT) one needs a 100 % increase of material (from 4.300 to 8.600 events; the full train set size)
- **Statistical Significance.** We refrained from testing the statistical significance of our results. The major issue we faced is the data sparseness of full texts (another strong point for more annotated full texts): to perform said test under minimal requirements, we would need to split the full text material in such a way that either the training or test set would hold too few events to be representative.
- **Error Analysis.** To provide a thorough error analysis, we would need to deal with 16 cases (all four systems with all four combinations of AB and FT for training and evaluation). We leave this discussion for a companion paper.

The outlier results for the HDS4NLP scores also deserve several remarks: The system

- achieves the highest F-Scores when trained and evaluated on abstracts—54.37; see Table 4 (**AB on AB**),
- yet has the lowest difference when these AB models are evaluated on full texts instead (Δ =–4.07); see Table 5, column 2,
- drops the highest (Δ = –8.59) when FT-trained instead of AB-trained models are used for evaluation on the AB test sets; see Table 5, column 3
- and further drops the highest when comparing **AB-AB** *vs.* **FT-FT** (Δ= –6.56); see Table 5, column 1.

We tend to explain this special role of the HDS4NLP system as a result of taking an entirely different methodological approach than the other three systems—HDS4NLP directly extracts pairwise structured events (and the event type they belong to) of the form (*trigger*, *argument*) instead of using a linear technique (i.e., extracting triggers first and then looking for applicable arguments). From the three competitive systems, HDS4NLP performs the worst in a solely **FT on FT** set-up so that these preliminary analyses could lead to the cautious conclusion that the HDS4NLP system is best (and better) suited to utilize abstracts as training material.

The exceptions coming from the JREX system (see its inconsistent increase in performance when Full Text models are evaluated on Abstract data, Table 5, column 4) are completely overshadowed by a substantial performance penalty in comparison with all other systems. This is clearly an effect of lacking maintenance over the past five years.

Conclusions

We focused in this paper on the implications of running established tooling and pipelines, developed on scientific abstracts as training data, on scholarly full text data from the life sciences. Our experiments reveal that F-score losses up until 6.6 F-Score points have to be anticipated. Abstracting away from the specific system particularities, there is a consistent trend for performance degradation when abstract-trained models are transferred without changes to full texts.

The apparent solution—providing annotated full-text corpora—is costly and resource-intensive. Currently, only one corpus (ID) provides relation encodings for full texts, whereas another one contains only highly specific named entity encodings (CRAFT). Given the resource-density in the life sciences field, distant supervision (via database contents) might be a reasonable alternative [30], while moving to unsupervised relation extraction is likely to be accompanied with an additional penalty in terms of sliding F-scores. Hence, approaches decomposing full texts into more digestible text portions (section and paragraph segmentation at the macro-level of text analysis; sentence simplification at the micro level) might be worthwhile to be taken into account.

References

[1] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, Genies: A natural language processing system for the extraction of molecular pathways from journal articles, *Bioinformatics* **17** (Suppl 1) (2001), S74–S82.

[2] P. Shah, C. Perez-Iratxeta, P. Bork, and M. Andrade, Information extraction from full text scientific articles: Where are the keywords? *BMC Bioinformatics* **4**(20) (2003).

[3] D. Corney, B. Buxton, W. Langdon, and D. Jones, BioRAT: Extracting biological information from full-length papers, *Bioinformatics* **20**(17) (2004), 3206–3213.

[4] L. K. Tanabe and W. J. Wilbur, Tagging gene and protein names in full text articles, in *BioNLP 2002 — Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain @ ACL 2002* (2002), 9–13.

[5] Z. Lu *et al.*, The Gene Normalization Task in BioCreative III, *BMC Bioinformatics* **12**(S8) (2011), S2.

[6] A. A. Morgan *et al.*, Overview of BioCreative II Gene Normalization, *Genome Biol* **9**(S2) (2008), S3.

[7] K. B. Cohen, H. L. Johnson, K. M. Verspoor, C. Roeder, and L. E. Hunter, The structural and content aspects of abstracts versus bodies of full text journal articles are different, *BMC Bioinformatics* **11** (2010), 492.

[8] C. O. Tudor, K. E. Ross, G. Li, K. Vijay-Shanker, C. H. Wu, and C. N. Arighi, Construction of phosphorylation interaction networks by text mining of full-length articles using the eFIP system, *Database – J. Biol. Databases Curation* #bav020, 2015.

[9] T. McIntosh and J. R. Curran, Challenges for automatically extracting molecular interactions from full-text articles, *BMC Bioinformatics* **10** (2009), 311.

[10] K. M. Verspoor *et al.*, A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools, *BMC Bioinformatics* **13** (2012), 207.

[11] M. Bada *et al.,* Concept annotation in the CRAFT corpus, *BMC Bioinformatics* **13** (2012), 161.

[12] S. Pyysalo *et al.*, Overview of the ID, EPI and REL tasks of BioNLP Shared Task 2011, *BMC Bioinformatics* **13** (S11) (2012), S2.

[13] J.-D. Kim *et al.,* The Genia Event and Protein Coreference tasks of the BioNLP Shared Task 2011, *BMC Bioinformatics* **13** (Suppl 11) (2012), S1.

[14] E. Buyko, E. Faessler, J. Wermter, and U. Hahn, Event extraction from trimmed dependency graphs, in *BioNLP – Proceedings of the Workshop on Biomedical Language Processing @ NAACL-HLT 2009. Shared Task on Event Extraction (*2009) 19–27.

[15] C.-H. Wei *et al.*, Assessing the state of the art in biomedical relation extraction: Overview of the BioCreative V chemical-disease relation (CDR) task, *Database – J Biol Databases Curation* baw032, 2016.

[16] C. N. Arighi *et al.*, BioCreative-IV virtual issue, *Database – J Biol Databases Curation*, bau039, 2014.

[17] C. H. Wu *et al.*, BioCreative-2012 virtual issue, *Database – J Biol Databases Curation*, bas049, 2012.

[18] C. Nédellec *et al.*, Overview of BioNLP Shared Task 2013, in *Proceedings of the BioNLP Shared Task 2013 Workshop (*2013), 1–7.

[19] J.-D. Kim, S. Pyysalo, T. Ohta, R. Bossy, N. Nguyen, and J.'i. Tsujii, Overview of BioNLP Shared Task 2011, in *Proceedings of BioNLP Shared Task 2011 Workshop*, (2011), 1–6.

[20] Ö. Uzuner and A. Stubbs, Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth Shared Tasks, *J Biomed Inform* **58** (S1–S5) (2015).

[21] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, 2010 i2b2/VA Challenge on concepts, assertions, and relations in clinical text, *JAMIA* **18**(5) (2011), 552–556.

[22] I. Segura-Bedmar, P. Martínez, and M. Herrero-Zazo, Lessons learnt from the DDIExtraction-2013 Shared Task, *J Biomed Inform* **51** (2014), 152–164.

[23] I. Segura-Bedmar, P. Martínez, and D. Sánchez-Cisneros, The 1st DDIExtraction-2011 Challenge Task: Extraction of drug-drug interactions from biomedical texts, in *DDIExtraction – Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction 2011. SEPLN 2011 Satellite Workshop* (2011), 1–9.

[24] J. Björne and T. Salakoski, Tees 2.1: Automated annotation scheme learning in the BioNLP 2013 Shared Task, in *Proceedings of the BioNLP Shared Task 2013 Workshop* (2013), 16–25.

[25] J. Björne, F. Ginter, and T. Salakoski, University of Turku in the BioNLP'11 Shared Task, *BMC Bioinformatics* **13**(11) (2012), 1–13.

[26] Q.-C. Bui, D. Campos, E. M. van Mulligen, and J. A. Kors, A fast rule-based approach for biomedical event extraction, in *Proceedings of the BioNLP Shared Task 2013 Workshop* (2013), 104–108.

[27] Q.-C. Bui and P. M. A. Sloot, A robust approach to extract biomedical events from literature, *Bioinformatics* **28**(20) (2012), 2654–2661.

[28] X. Liu, A. Bordes, and Y. Grandvalet, Biomedical event extraction by multi-class classification of pairs of text entities, in *Proceedings of the BioNLP Shared Task 2013 Workshop (*2013), 45–49.

[29] E. Buyko, E. Faessler, J. Wermter, and U. Hahn, Syntactic simplification and semantic enrichment: Trimming dependency graphs for event extraction, *Computational Intelligence* **27**(4) (2011), 610–644.

[30] E. K. Mallory, C. Zhang, C. Ré, and R. B. Altman, Large-scale extraction of gene interactions from full-text literature using DeepDive, *Bioinformatics* **32**(1) (2016), 106–113.

Address for correspondence

Franz Matthies: franz.matthies@uni-jena.de