On the Rule-based Extraction of Statistics Reported in Scientific Papers

Tobias Kalmbach, Marcel Hoffman[0000−0001−8061−9396], Nicolas Lell[0000−0002−6079−6480], and Ansgar Scherp[0000−0002−2653−9245]

Universität Ulm, Germany `firstname.lastname@uni-ulm.de`

Abstract. The identification and extraction of statistics in scientific papers as nested entities is an indispensable feature for analyzing scientific papers at a large scale. STEREO is a tool for extracting statistics from scientific papers using a set of regular expressions. Key feature of the tool is that it supports statistics reported in American Psychology Association (APA) style, as well as non-APA style such as only a reported *p*-value. The original STEREO rule set has been extensively trained in the life sciences domain using preprints of the CORD-19 dataset. We analyze this rule set with its hundreds of regular expressions using a regular expression inclusion algorithm. We transfer the condensed rule set to papers in the domain of Human-Computer-Interaction (HCI). Our experiments show that only 13 new R^+ rules and 77 new R^- rules are needed to conduct this transfer. A higher percentage of APA-conform statistics were found in the HCI domain (26%) compared to the life sciences domain (only 1.8%). We compare the statistics extraction from PDFs vs. LaTeX source files, finding the latter more reliable.
An extended version with detailed examples is provided on arXiv [11] and the source code is here: `https://github.com/Tobi2K/statistics-extraction`

Keywords: statistics extraction · nested entities · regular expressions.

1 Introduction

An abundance of scientific papers are published daily. The large and rapidly growing number of papers is too extensive to scan manually. In particular, assessing the published results in terms of insights generated by the statistical analyses such as significance tests is very challenging. A quick overview of the statistics in a paper can also be useful for the authors to find statistical errors in their studies, i. e., to verify and check them. Moreover, extracting sentences containing statistics together with metadata (authors, title, etc.) can enable researchers to get an impression of an article without the need to read it. Tools like *statcheck* [16] provide very accurate extraction of statistics reported in accordance with the commonly used writing style guide of the American Psychology Association (APA) [2]. However, previous research found less than one percent of APA-conform statistics in a sample of 113, 000 statistics extracted from preprints in the life sciences [7]. In this work, we extend STEREO (STat ExtRaction Experimental cOnditions) [7], an automatic statistics extraction pipeline for statistics presented in APA as well as non-APA notation.

STEREO learns regular expressions (rules) to decide whether a sentence contains statistics (R^+ rules) or not (R^- rules) using active wrapper learning. The R^+ rules are used to extract the statistic's type and values. During the application of STEREO on the life sciences dataset CORD-19 [20] containing preprints of papers about the corona virus and related viruses, a total of 85 R^+ rules for statistics detection (with 52 sub-rules for value extraction) and $1,425$ R^- rules were learned. Inspecting these rules shows that rules, which were added later in the learning process, generalize better and previously added rules become obsolete.

Reducing the number of rules can help to identify common patterns of non-APA statistics, e. g., incomplete reporting, and derive recommendations to improve statistics reporting. Subsequently, general rule patterns that indicate a sentence does not contain a statistic can serve as guidance for future active wrapper approaches. One can make use of these general rule patterns to avoid creating excess rules. Following this reasoning, we apply a DFA-based (Deterministic Finite Automaton) algorithm introduced by Chen and Xu [5] to minimize the existing set of regular expressions.

We transfer STEREO to a new scientific domain, namely Human-Computer-Interaction (HCI), to investigate the generalization of the STEREO rules and potentially find new rules for statistics extraction. This includes finding uncovered statistic types and other non-APA conform reporting of statistics. We further extend the STEREO rules from text to also support statistic extraction from LaTeX files. In summary, this work makes the following contributions:

- We analyze the extraction rules from STEREO and achieve a rule set reduction of 34%, which results in 31% less runtime needed to apply the reduced rule set compared to the full rule set.
- We extend the rule set by repeating the active wrapper learning from STEREO on the HCI domain, adding 13 new R^+ rules and 77 R^- rules.
- Using the new rule set, we identify that 26% of all statistics extracted from HCI preprints are in APA style, while in the life sciences domain we found only 1.8% of statistics to be conform to APA.
- We compare the extraction from LaTeX versus PDF files. The extraction precision is high in both cases. However, we miss 20% of the statistics in PDF due to transformation errors from PDF to text.

Below we discuss related work on statistics extraction and regular expression inclusion. Section 3 presents the experimental apparatus. The results are reported in Section 4 and discussed in Section 5, before we conclude.

2 Related Work

Statistics Extraction Statistic extraction poses several challenges like different writing styles or usage of number separators and might even require to parse formulaic expressions [9]. Teja et al. [12] presented a regular expression-based approach to extract statistics from scientific papers. They use a single regular expression to match the p-value per statistical test. For example, a regular expression to match a t-test is `t(df)=float, p (<, >, =) float`. A similar approach is pursued by *statcheck* [16], an R package that allows to extract and verify the consistency of statistics reported following APA guidelines. If all information required by APA is provided, e. g., the p-value and degrees of freedom, statcheck can check if the reported statistics is plausible. Recomputing the statistics is not possible since this would require access to the raw data. Schmidt [19] disputed the effectiveness of *statcheck*. They criticize the testing conducted in the *statcheck* authors' follow-up paper [15]. Schmidt [19] argues that *statcheck* simply does not detect many reported statistics due to the strong assumption that they must be reported following APA. This questions the overall performance of *statcheck*, even though it is widely used [8,18,17].

The approaches by both Nuijten et al. [16] and Teja et al. [12] are limited to only match APA-style statistics. Böschen [3] presented a text-mining approach on XML documents. They differentiate between computable results, where the p-value is given and can be recalculated; checkable results, where the p-value is not given but can be calculated; and uncomputable results, where the p-value cannot be calculated due to some information missing. The extraction algorithm by Böschen [3] works as follows: Sentences are only selected if they contain at least one letter, followed by an operator ($<, >, =, \leq, \geq$), which in turn is followed by a number. Surrounding text is removed using regular expressions. Individual heuristics are applied to extract the recognized test statistics, the operator, degrees of freedom, and p-value to cope with varying reporting styles. As the requirements are not as strict as *statcheck*'s, Böschen [3] generally finds more statistics. STEREO [7] uses active wrapper learning to learn regular expressions (rules) that determine whether or not a sentence contains statistics. The rules are divided into R^+ rules that match statistics and R^- rules that denote that a sentence does not include statistics. R^+ rules have additional sub-rules, which are used to capture specific parts of the statistic (e. g., the p-value) after the statistic type has been identified by the main R^+ rule. During the active wrapper learning, every sentence that does not contain a number is ignored. For any remaining sentences not matched by any rules, the user is prompted to create a new rule to cover the new case. STEREO achieved a precision close to 100% for APA-conform statistics and 95% for non-APA statistics on the CORD-19 dataset.

Rule Set Inclusion Algorithms Regarding regular expression inclusion, i. e., minimal rule set computation, many algorithms are limited to determining inclusion using one-unambiguous regular expressions. One-unambiguous regular expressions [4] are a subset of regular expressions that can match every word (in their respective language) in a unique way without looking ahead. For example, $(a_1|b_1)^*(a_2|\epsilon)$ (numbered for clarity) is not one-unambiguous, as the word *baa* can be formed as $b_1a_1a_1$ or $b_1a_1a_2$. However, $(a|b)^*$ describes the same language but is one-unambiguous. Chen and Xu [5] presented two algorithms for regular expression inclusion. The first is an automata-based algorithm that converts the given one-unambiguous regular expressions into Deterministic Finite Automatons (DFAs) and subsequently checks the created DFAs for inclusion. The second algorithm is a derivative-based algorithm. Derivatives of regular expressions are sub-expressions, which are valid regular expressions themselves. The idea is that if an expression A is included in an expression B, all derivatives of A are also derivatives of B. The algorithm generates all derivatives of both expressions. If all derivatives of one expression are included in the other, the first expression is included in the second. Nipkow and Traytel [14] presented a framework to determine if two given regular expressions are equivalent. Equivalence is a stricter requirement than the inclusion of Chen and Xu [5]. The framework dynamically creates an automata from one regular expression and uses "computations on regular expression-like objects" [14, p. 2] as a substitute for the traditional transition table. Hovland [10] presented an approach that uses an inference system instead of automata to inductively determine a binary relationship between one-unambiguous regular expressions. The algorithm guarantees polynomial runtime, which can be slower than the quadratic runtime of Chen and Xu [5].

3 Experimental Apparatus

Datasets We have two types of datasets: The original rule set from STEREO and the scientific papers in life sciences and the new HCI domain.

STEREO's rule set for the life sciences was created using the COVID-19 Open Research Dataset (CORD-19). We apply minimal rule set analysis on this dataset. The dataset consists of $1,510$ manually created rules, divided into 85 R^+ and $1,425$ R^- rules. Each rule has an incremental ID (determined at the time of creation) and its corresponding regular expression. Implicitly, a higher rule ID means that the rule has been added later in the process of applying the active wrapper. We assume that it is unlikely that the 85 R^+ rules can be optimized greatly, as these rules are designed to match specific information present in the reporting of a statistics. This makes it unlikely that one R^+ rule is included in another. However, the $1,425$ R^- rules can be optimized to improve runtime performance as well as maintainability by revealing common patterns used to identify sentences as non-statistic.

The pre-print papers in life sciences and HCI: The COVID-19 Open Research Dataset [20] is the original dataset used in STEREO that can be used to evaluate the minimal rule set. This dataset contains $110,427$ papers provided in JSON-format on COVID-19, SARS-CoV-2, and all corona viruses in general. In STEREO, the date of access is given as 21st September 2020. The CORD-19 dataset version 52[1], which we use for comparison, is a close match. Note, that our version of the dataset is slightly newer and contains a few more papers than

[1] (publication: 2020-09-21, accessed: 2022-07-11) https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/versions/52

the version used in STEREO. Thus, we rerun the experiments of the original paper for a fair analysis of the rule set.

The arXiv Dataset [6] has over 1.7 million STEM papers.[2] It includes metadata like author, category, etc. We filter for HCI, which studies the use of technology, focusing on the interface between people and computers [13]. HCI is a strong domain for publishing studies and the corresponding statistics. There are $9,730$ papers tagged with the "cs.HC" (HCI tag on arxiv.org) category. We only use papers with HCI as primary tag. For a fair comparison of the statistics extraction on PDF and LaTeX, we only use papers that are provided in both formats. With these restrictions, $4,023$ papers remain.

Preprocessing For the rule set inclusion, we transform the regular expressions provided by STEREO from the original Python format into a formal representation that is required for the inclusion algorithm of Chen and Xu [5]. For the transfer of STEREO's rules from the life sciences to HCI, we parse the content to plain text while removing all `table`, `figure`, `lstlisting`, and `tikzpicture` environments. As in STEREO, line breaks are removed and the plain text is split into sentences using the regular expression `\.\s?[A-Z]`. Every sentence that does not contain a digit is removed. The corresponding PDF files are converted to raw text using *pdftotext*[3]. Then the same processing (line breaks, split sentences, keep numbers) is applied. This results in $9,393,662$ sentences for the CORD-19 dataset and $222,544$ sentences for the HCI domain.

Procedure We compute the minimal set of R^+ and R^- rules on the STEREO rule set for life sciences using the M_{E_1}-directed version of the algorithm from Chen and Xu [5]. Although the runtime of this algorithm is quadratic in length of the expressions, it is sufficient as our regular expressions are usually short (< 100 characters) resulting in a good trade-off between runtime and effort to implement it. We do a pairwise comparison of the rules in the set. A rule that is already covered by some other rule is removed.

We randomly sample two times 200 papers (about 5% of all papers) from the $4,023$ papers in the HCI dataset, one sample contains LaTeX the other PDF files. On these samples, we repeat the active wrapper induction from STEREO to learn new rules for the extraction of statistics from HCI papers. Specifically, we split the input into sentences and STEREO checks for every digit in the sentence if it is covered by an existing rule. In case of any uncovered digit, the active learning approach of STEREO prompts a user interface and asks an expert to add a new R^+ or R^- rule, based on whether the sentence contains statistic or not. Each new rule is assigned an incremental ID. This step produces our new rules that are added to the STEREO rule set. As we use new input formats, i. e., LaTeX versus PDF, this results in new, format-specific rules. The goal is to further improve the robustness and completeness of the extraction.

[2] https://www.kaggle.com/datasets/Cornell-University/arxiv (2022-07-11)
[3] https://pypi.org/project/pdftotext/

Finally, we evaluate the precision for every statistic type following the evaluation procedure of STEREO [7]. We extract all sentences from all papers in the respective corpus that are matched by R^+ rules. We then sample 200 sentences for every statistic type, or use all extracted sentences if there were fewer than 200 extractions, to manually check if the statistic types are matched and extracted correctly. We also measure the difference in APA versus non-APA reporting. Furthermore, we extract 200 sentences matched by R^- rules from random documents, which we did not use for rule learning. This is to test whether R^- rules do not reject statistics, i. e., we have false negatives. Lastly, we extract 200 sentences that were neither matched by any R^+ nor R^- rule to check for unrecognized statistics or data format transformation errors, i. e., errors that were introduced when transforming a paper from the input format (e. g., PDF) to plain text [7].

Measures For the rule set inclusion, we measure runtime and number of included rules per rule. Furthermore, we check that the reduced and the original rule set cover the same sentences. Finally, we compare the time required to match 200 sentences with the original rule set versus the reduced R^- rule set.

For the experiments on comparing statistics extraction in HCI versus life sciences and from LaTeX versus PDF files, we use precision of the extraction. We calculate it on 200 rules per statistic type or the maximum amount when there are less than 200 extractions.

4 Results

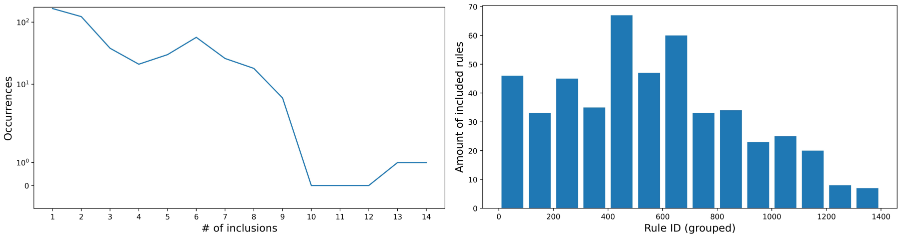
Rule Set Inclusion After running the rule inclusion algorithm for the R^- rules, 483 unique rules out of the total $1,426$ rules were included by others. This is a reduction of 33.8%. $1,253$ rules included no other rules, 83 included one, and 28 included two rules. However, the analysis also revealed that 13 rules included more than 20 rules, and 4 had more than 100 rules included. These rules can be seen in Table 1. Naturally, some rules were included more than once. Figure 1a shows how often rules were included in other rules. The rule `figure \d{1,2}` was the most included one with 14 inclusions, and `table \d+` was the second-most included rule with 13 inclusions.

We show the included rules sorted by rule ID (grouped in hundreds) in Figure 1b. The ID of each rule reflects when it was created in the active wrapper process, i. e., rules with lower IDs were created earlier than rules with higher IDs. We observe that many rules that were included by others had a rule ID between 400 and 700. Furthermore, approximately 47% of included rules had an ID below 500 and 88% had an ID below $1,000$. In general, lower ID rules are included in higher ID rules.

We also ran the inclusion algorithm on the R^+ rules to double check whether optimization is possible. We observe that one rule was removed. This inclusion was an exact duplicate that most likely was added by mistake. Thus, the R^+ rules do not need further consideration for rule set minimization.

Table 1: R^- rules sorted in descending order by the number of included rules and having more than 50 inclusions.

Regular Expression	# included rules
`[a-zA-Z]{3,}\s?\d+[\.\,\s\dabcdef]*`	173
`[a-zA-Z]{2,}\s?\d+(\.\d)?`	173
`[a-zA-Z]{3,}\s?\d+[\.\,\s\d]*`	171
`[a-zA-Z]{3,};?\s?\d+`	130
`[a-zA-Z"]+\s?\d{1,3}$`	83
`[a-zA-Z]{3,20}\s\d+(\,\d+)*(\.\d+)?`	72
`[a-zA-Z]{3,20}\d+`	71
`[a-zA-Z]{3,}\s-?\d+(\.\d+)?`	62
`\d+(\,\d+)*(\.\d+)?\s[a-zA-Z]{3,10}`	51



(a) Number of times a rule was included by other rules.

(b) Amount of rules included in other rules, sorted by ID, grouped by hundreds.

Fig. 1: Left: 166 rules were included once, one rule was included 14 times. Right: Amount of inclusions based on rule ID. The Rule ID reflects the order in which a rule was created in STEREO's [7] wrapper induction (see Sec. 3).

Transfer of the Rules to the HCI Domain Using the LaTeX files, we added 13 new R^+ rules and 77 R^- rules. Furthermore, we manually changed 6 previously added R^- rules to be more general. For example changing `m^2` to `m^[2|3]` to capture both square and cubic meters. The R^+ rules added two new statistics types, the Z-Test and ANOVA without an r-value. The statistics covered by our rules are those frequently used in HCI literature [13]. In the original implementation of STEREO, all ANOVA tests that did not contain a r-value were seen as non-APA. However, we found that APA guidelines do allow ANOVA to be reported without an r-value [1]. Therefore, when referencing the percentage of APA-conform statistics in a corpus, we mention both including and excluding ANOVA tests without an r-value. For both the Z-Test and ANOVA without an r-value, only APA conform extraction rules were added.

For the PDF files, nine R^- rules and no R^+ rule had to be added. These new R^- rules where added because the PDF to text conversion includes page numbers and citations, which are not contained in the LaTeX files of the HCI dataset and the JSON files of the CORD-19 dataset.

Table 2: Number (*num*) of extracted statistics and precision (P) over $\min(200, num)$ for APA and non-APA conform reporting on HCI papers. Separately considering the extraction from PDF and LaTeX files.

	APA conform				non-APA conform			
	PDF		LaTeX		PDF		LaTeX	
Statistic Type	*num*	P	*num*	P	*num*	P	*num*	P
Student's t-test	440	100%	634	100%	38	97.4%	69	100%
Pearson Correlation	48	100%	65	100%	76	96%	94	96.8%
Spearman Correlation	2	100%	1	100%	59	90.7%	64	89%
ANOVA	0	N/A	0	N/A	2	100%	0	N/A
ANOVA without r-value	1,059	100%	1,097	100%	0	N/A	0	N/A
Mann-Whitney-U	0	N/A	0	N/A	270	92%	425	94%
Wilcoxon Signed-Rank	0	N/A	0	N/A	0	N/A	0	N/A
Chi-Square	53	100%	85	100%	14	100%	718	100%
Z-Test	66	100%	195	100%	0	N/A	0	N/A
Total supported statistics	1,668		2,077		459		1,370	

In total, transferring STEREO from the life sciences to the HCI domain required adding 99 rules, including 13 R^+ rules to cover Z-Tests and ANOVA without r-value.

Precision of Statistics Extraction for the HCI Dataset Our R^+ rules extract the same statistic types used in STEREO [7] (Pearson's Correlation, Spearman Correlation, Student's t-test, ANOVA, Mann-Whitney-U Test, Wilcoxon Signed-Rank Test, and Chi-Square Test). These statistic types were chosen as they were commonly found in scientific papers. We added two new types of statistics found often in HCI papers, the Z-Test and ANOVA without an r-value.

In the 4,023 HCI papers, the R^+ rules matched 6,321 sentences from the PDF files and 7,669 sentences from the LaTeX files. Normalizing this to the total amount of sentences in both file types, these numbers correspond to about 3% of the sentences. Table 2 shows all reported statistics categorized by type and whether the statistics matched APA style or not. For every statistic type, more statistics were extracted from LaTeX files than from PDF files. The only exception is the Spearman Correlation in the case of APA conform and LaTeX. We denote 'Other Statistics' as all statistic types which are extracted by STEREO but are not assigned a specific APA or non-APA type. It includes a range of statistics not yet captured by the rule set, e. g., interquartile range or Kolmogorov-Smirnov tests. We do not list 'Other Statistics' in the result tables. Using PDF files, about 26% of the extracted statistics were APA conform (9% when treating ANOVA without r-value as non-APA). With LaTeX files, 27% of extracted statistics were APA conform (13% when considering ANOVA without r-value as non-APA).

On PDF files, the precision for APA statistics was 100% and ranged from 90% to 100% for non-APA statistics (see Table 2). 'Other Statistics' had a precision of 54.5% with 4,194 extracted statistics. Similarly, using the LaTeX files, we achieved 100% precision for APA conform statistics and precision ranging from 89% to 100%, otherwise. 'Other Statistics' had an increased precision of 60.5% but only 4,184 extracted statistics. Adding the 'Other Statistics', we extracted 1,668 APA conform and 4,653 non-APA conform statistics on PDF files. On the LaTeX files, we extracted 2,077 APA conform and 5,554 non-APA conform statistics in total.

Precision of the Statistics Extraction for the CORD-19 Dataset The number of statistics extractions and their precision on the CORD-19 dataset are presented in Table 3. As expected, we achieve similar results as the original STEREO paper. Please note that, as mentioned earlier, STEREO used a slightly older version of the dataset than the one available to us. The most statistic was extracted for 'Other Statistics' with 114,242 and a precision of 98.5%. In total, 2,189 APA conform and 120,516 non-APA conform statistics were extracted. For the supported statistic types, non-APA conform Pearson Correlations were extracted the most, by a large margin. Of the extracted statistics, 1.8% were APA conform (0.8% when treating ANOVA without r-value as non-APA). As for the HCI dataset, the APA-conform extractions achieved a precision of 100%. For non-APA conform statistics, the precision ranged from 94.5% to 100%.

Table 3: Number (*num*) of extracted statistics for APA and non-APA conform reporting on CORD-19 papers. Precision (P) is calculated on 200 samples per type or all samples if there are less.

	APA conform		non-APA conform	
Statistic Type	*num*	P	*num*	P
Student's t-test	662	100%	210	97%
Pearson Correlation	113	100%	5,034	98.5%
Spearman Correlation	1	100%	551	100%
ANOVA	0	N/A	2	100%
ANOVA without r-value	1,239	100%	0	N/A
Mann-Whitney-U	2	100%	419	94.5%
Wilcoxon Signed-Rank	0	N/A	0	N/A
Chi-Square	69	100%	58	100%
Z-Test	103	100%	0	N/A
Total supported statistics	2,189		6,274	

R^- rule evaluation We evaluate the R^- rules on 200 randomly selected sentences from the HCI dataset as well as the CORD-19 dataset. We aim to test if any reported statistic was falsely matched by R^- rules, i. e., results in a false negative. Our investigation shows that for both datasets, all 200 sentences were correctly identified as non-statistics. We measure the runtime using the HCI dataset and compare the reduced rule set with the full rule set. The full rule set takes 122.4 seconds (averaged over 5 runs), while the reduced rule set takes 84.5 seconds

(averaged over 5 runs). This is a performance gain of about 31% for the reduced rule set.

We extract 200 sentences that contain a number but were not matched by any R^- or R^+ rules. We assess whether these uncaptured sentences report a statistic or not, or whether they contain a transformation error regardless of the content (see Table 4). 92% of uncaptured sentences did not contain any statistics. The most missed statistics (8.5%) were in CORD-19, whereas using the PDF files in the HCI domain missed the fewest (2.5%). Using the CORD-19 dataset resulted in the most transformation errors, while using LaTeX did not have transformation errors.

Table 4: Evaluation of sentences not covered by R^- or R^+ rules. Evaluated on a sample of 200 sentences taken from the respective datasets.

Dataset	Statistic missed	No statistic contained	Transformation error
CORD-19 + JSON	17 (9%)	174 (87%)	9 (5%)
HCI + LaTeX	14 (7%)	186 (93%)	0 (0%)
HCI + PDF	5 (3%)	192 (96%)	3 (2%)

5 Discussion

Statistics Extraction from the Datasets Using the HCI dataset, about 26% of the extracted statistics were APA conform. This is a large difference to the 1.8% of APA conform statistics in the CORD-19 dataset. Nonetheless, this means that the remaining 74% for HCI and 98.2% for CORD-19 of reported statistics are non-APA conform. This makes understanding the scientific progress and relying on studies very difficult for researchers, as discussed in the introduction. Since all APA-conform statistics follow a very strict and well-defined pattern, they achieve a precision of 100%. However, non-APA Mann-Whitney-U test rules need refinement, as in all scenarios, some Wilcoxon Signed-Rank tests were falsely identified as Mann-Whitney-U tests.

Generally, we could extract more statistics from LaTeX files than using PDF files in the HCI dataset. Note that the dataset contains only preprints that are available in both formats. This means we loose statistics in the process of converting PDF to text and learning rules to find those statistics in the text. However, it is encouraging that regardless of the file format, the precision of the extracted statistics is generally high for both PDF and LaTeX.

In LaTeX files and in the CORD-19 dataset, page numbers, as well as citations, were automatically removed or never generated. However, converted PDF files contained citations, which in turn included pages of an article in a journal or ACM identifiers. These had a high diversity of representation, which makes defining new R^- rules to capture them very difficult. Some examples of these variations can be seen in the following:

- `Human Factors in Computing Systems. dl.acm.org, 2853{2859.`
- `ACM, New York, NY, USA, 285{296.`
- `Computer Graphics 19, 12 (2013), 2713{2722.`
- `Virtual Environ. 7(3), 225{ 240 (1998).`
- `Thousand Oaks, CA, 508{510 (2007) [49]`

In the end, we added the rule `\),\s\d{1,4}[-{]\d{2,4}[.)]` to capture most cases. Tables could not be removed from the PDF input, leading to some extra rules. However, most numbers were already matched by the previously added R^- and R^+ rules.

We performed a detailed analysis on the large deviation of the precision (Table 2) for 'Other Statistics', which are statistics we extract but do not determine the type of, compared to explicitly captured statistic types, i. e., those that are supported by specific and typed R^+ rules. We identified a rule in STEREO that is `\([P|p] \s? <?=? \s? \d (\.\d+)?\)`. This rule also captures the string `(P1)`, which is not a statistic and produced false positive. Thus, we change the rule to `\([P|p] \s? [<=]+ \s? \d (\.\d+)?\)`. We re-run the evaluation and retrieve $2,337$ 'Other Statistics'. Now the precision goes up to 97.5% for the PDF files. For the LaTeX files, 'Other Statistics' extractions are reduced to $2,254$, with a precision increase to 98.5%.

Inspecting the Reduced Set of R^- Rules We applied the rule set inclusion algorithm from Chen and Xu [5] to reduce the rule set of STEREO. The goal is to improve STEREO's runtime, which we observe to be by about one third. A detailed inspection on the rule inclusion in Figure 1b shows that later added rules are more likely to include one or more other rules. Note, "later" refers to the point in time a rule was added during the active wrapper learning process, i. e., a higher rule ID was assigned to it (see Section 3). We assume that later rules were added with more background knowledge of the STEREO tool and thus they tend to be more general. The most included rule is `figure \d{1,2}`. Later rules like `(...| fig | figure | Table |...)\s*\d+(\s*[\.\,]\s*\d+)*` (shortened) include the first rule and do not only match a figure, but also tables and equations.

The structure of rules with many inclusions is mostly similar. Every rule, which included more than 100 other rules, leveraged numbers being preceded or followed by a word. Fore example, `[a-zA-Z]{3,}` covers rules of the same structure designed for special physical units like `\d[mM]` matching meter information.

6 Conclusion

We analyzed STEREO, which extracts statistics from papers using a set of regular expressions. We apply a rule set inclusion algorithm that removed a third of the rules. We extend the rule set to the HCI domain. We repeated the active wrapper learning from STEREO on a sample of 200 papers, i. e., $222,544$ sentences. We only had to add 13 R^+ and 77 R^- rules to cover this new domain. This is a small fraction of newly required rules compared to the $1,510$ original rules in STEREO. We apply the extended statistics extraction rule set to the whole HCI dataset. We find that only 26% of extracted statistics were APA conform in the HCI domain, compared to only 1.8% for the CORD-19 dataset.

We compare the use of PDF versus LaTeX files in the HCI domain. The overall extraction precision is high independent of the format. For PDF converted to text, we observe a few transformation errors, which do not occur with LaTeX.

In future studies, one could further analyze Wilcoxon Signed-Rank tests that were often falsely captured as Mann-Whitney-U tests. While Wilcoxon Signed-Rank and Mann-Whitney-U have similar reporting styles, exploiting more surrounding context might better separate these two types.

Acknowledgements

This work is co-funded under the 2LIKE project by the German Federal Ministry of Education and Research (BMBF) and the Ministry of Science, Research and the Arts Baden-Württemberg within the funding line Artificial Intelligence in Higher Education.

References

1. APA: Publication manual of the American Psychological Association 2020: the official guide to APA style. American Psychological Association, 7 edn. (2020)
2. Bentley, M., Peerenboom, C., Hodge, F., Passano, E.B., Warren, H., Washburn, M.: Instructions in regard to preparation of manuscript. Psyc. Bulletin (1929)
3. Böschen, I.: Evaluation of JATSdecoder as an automated text extraction tool for statistical results in scientific reports. Scientific Reports (2021)
4. Brüggemann-Klein, A., Wood, D.: One-unambiguous regular languages. Inf. Comput. (1998)
5. Chen, H., Xu, Z.: Inclusion algorithms for one-unambiguous regular expressions and their applications. Sci. Comput. Program. (2020)
6. Clement, C.B., Bierbaum, M., O'Keeffe, K.P., Alemi, A.A.: On the use of ArXiv as a dataset (2019), https://arxiv.org/abs/1905.00075
7. Epp, S., Hoffmann, M., Lell, N., Mohr, M., Scherp, A.: STEREO: A pipeline for extracting experiment statistics, conditions, and topics from scientific papers. In: iiWAS. ACM (2021)
8. Freedman, L.P., Venugopalan, G., Wisman, R.: Reproducibility2020: Progress and priorities. F1000Research (2017)
9. Göpfert, J., Kuckertz, P., Weinand, J., Kotzur, L., Stolten, D.: Measurement extraction with natural language processing: A review. In: EMNLP. ACL (Dec 2022)
10. Hovland, D.: The inclusion problem for regular expressions. J. Comput. Syst. Sci. (2012)
11. Kalmbach, T., Hoffmann, M., Lell, N., Scherp, A.: Reducing a set of regular expressions and analyzing differences of domain-specific statistic reporting. CoRR **abs/2211.13632** (2022), https://arxiv.org/pdf/2211.13632v2.pdf
12. Lanka, S.S.T., Rajtmajer, S.M., Wu, J., Giles, C.L.: Extraction and evaluation of statistical information from social and behavioral science papers. In: Companion of The Web Conference 2021. ACM / IW3C2 (2021)

13. Lazar, J., Feng, J., Hochheiser, H.: Research Methods in Human-Computer Interaction. Morgan Kaufmann (2017)
14. Nipkow, T., Traytel, D.: Unified decision procedures for regular expression equivalence. In: Interactive Theorem Proving - 5th International Conference, ITP 2014, Proceedings. Lecture Notes in Computer Science, Springer (2014)
15. Nuijten, M.B., van Assen, M.A.L.M., Hartgerink, C.H.J., Epskamp, S., Wicherts, J.M.: The validity of the tool "statcheck" in discovering statistical reporting inconsistencies (2017), psyarxiv.com/tcxaj
16. Nuijten, M.B., Hartgerink, C.H., Van Assen, M.A., Epskamp, S., Wicherts, J.M.: The prevalence of statistical reporting errors in psychology (1985–2013). Behavior research methods (2016)
17. PsychOpen: Psychopen uses statcheck tool for quality check. PsychOpen (2017)
18. Sakaluk, J.K., Graham, C.A.: Promoting transparent reporting of conflicts of interests and statistical analyses at the journal of sex research. J. of Sex Research (2018)
19. Schmidt, T.: Statcheck does not work: All the numbers. reply to Nuijten et al. (2017) (2017), psyarxiv.com/hr6qy
20. Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., et al.: CORD-19: The Covid-19 Open Research Dataset. CoRR **abs/2004.10706** (2020)