Human-in-the-Loop Information Extraction Increases Efficiency and Trust

Johannes Schleith
Thomson Reuters Labs
London, United Kingdom

Hella-Franziska Hoffmann
Logically UK[1]
London, United Kingdom

Milda Norkute
Thomson Reuters Labs
Zug, Switzerland

Brian Cechmanek
Thomson Reuters Labs
London, United Kingdom

ABSTRACT

Automation is often focused on data-centred measures of success, such as accuracy of the automation or efficiency gain of individual automated steps. This case study shows how a human-assisted information extraction system, that keeps the human in the loop throughout the creation of information extraction rules and their application, can outperform less transparent information extraction systems in terms of overall end-to-end time-on-task and perceived trust. We argue that the time gained through automation can be wiped out by the perceived need of end users to review and comprehend results, where the systems seem obscure to them.

KEY WORDS AND PHRASES

Information Extraction, Human In The Loop, Human Centered AI

1 INTRODUCTION

Knowledge workers, who need to review information, lack easy-to-use, customizable tools to identify and extract specific entities from unstructured documents [1]. In this study we designed and evaluated a human in the loop information extraction system, that enables the user to express information extraction rules through keyword search and Named Entity Recognition (NER), and guided by Machine Learning (ML) to suggest improvements to such rules.

We studied a use case in information extraction for news production, which requires the extraction of highly specific, nuanced information. Some of the main challenges here are a "context gap", i.e. inability to convey search context, a lack of training data for full automation and a potential lack of trust into black box ML systems. We present an extraction system that keeps the human in the loop throughout extraction rule creation, and gradual improvement based on automated suggestions. We show that a human in the loop information extraction system outperforms less transparent extraction systems in terms of efficiency and trust. The study builds on other similar approaches [2][3], where information extraction rules were generated based on user specified examples.

[1] The research was conducted while being employed with Thomson Reuters Labs.

Veröffentlicht durch die Gesellschaft für Informatik e.V.
in K. Marky, U. Grünefeld & T. Kosch (Hrsg.):
Mensch und Computer 2022 – Workshopband, 04.-07. September 2022, Darmstadt

https://doi.org/10.18420/muc2022-mci-ws12-249

2 BACKGROUND

Thanks to recent advances in artificial intelligence (AI) and machine learning (ML), AI solutions are being built and integrated into many technology solutions across various sectors. However, balancing powerful ML capabilities with the need for end users to trust the system and to feel empowered can be challenging. Specifically, a sufficient level of perceived understanding of the technology and perceived control is important for end user trust, and ultimately adoption of the system [4][5].

Increased attention has been given to fair, interpretable, accountable and transparent algorithms in the AI and ML communities [6]. In 2016, the European Union approved a data protection law known as the General Data Protection Regulation or "GDPR" [7] that includes a "right to explanation". AI principles of various organizations state the need for decisions made by AI to be "explainable", "understandable" and "subject to human direction and control" [8] [9] [10]. AI practitioners look for concrete ways to explain the decisions made by their AI models in different contexts and use cases.

One such use case is information extraction (IE), the automated retrieval of specific information related to a selected topic from a body or bodies of text [11]. IE tools make it possible to pull information from text documents, databases, websites or other sources. However, most ML approaches to IE require large collections of labelled datasets for training that can be difficult to acquire or create. Mathematical models that are used in ML, often seem obscure to end users, occasionally returning unwanted results that are unexplainable [2]. Meanwhile, rule-based approaches can result in understandable and explainable IE rules, but might be time-consuming and labour-intensive to create [12].

Hannafi et al. [2] attempted to solve this by creating SEER, an IE system that suggests easy-to understand extraction rules from a small set of extracted-text examples provided by end users [2]. Users provide examples by highlighting text from documents that they wish to extract. Based on the highlighted examples, SEER learns extraction rules in Visual Annotation Query Language (VAQL) [13]. Similarly, Kejrival et al. [3] created a GUI-based editor to enable domain experts to construct knowledge graphs by writing sophisticated rule-based entity extractors with minimal training. Each

rule was implemented as a collection of user-defined SpaCy token sequence (e.g. "a number" followed by "character '£' or word 'Pounds'"). Building exhaustive token-pattern rules is time consuming and error-prone. Hence, many industry-grade IE solutions provide out-of-the-box extractors (e.g. spaCy[1]) for well-known text patterns like numbers, date, and monetary values via Regular Expressions (RegEx), and named entities like location and organization in form of Named Entity Recognition (NER) models.

Building on the above, we developed a GUI based system to both create and review rules for information extraction. It enables end users to customize such rules with the help of RegEx and NER-based suggestions.

3 USE CASE

Prior to planning the experiment, we conducted user research using interviews and contextual inquiries in news gathering, which showed that business and financial news reporting builds on high speed detection and reporting of unexpected events and key business figures. Reporting teams need to rapidly skim company reports for information on companies' revenue values, earnings per share, dividends, any potential increase or decrease of such numbers, new mergers and acquisitions, or change of a management of a company.

3.1 Lack of training data

On the one hand, such a high stakes use case requires extracted information to be absolutely correct, as it might impact further investment and business decisions. On the other hand, the need for unique and highly specific information, might limit the amount of training data that could be used for ML based automation, very similar to the "cold start" problem faced by many recommender systems [14]. This study proposes a concept that enables information extraction without any initial training data while maintaining the ability to turn user annotations into labelled training data for future ML based iterations of the system.

3.2 Context Gap

While off-the-shelf NER techniques help identify types of information (e.g. monetary values, dates, names), they might yield a large number of false positives (e.g. display of any monetary value rather than only earnings per share values). When using information extraction systems that are not customized or adapted to a specific use case, end users experience a "context gap", i.e. an inability to convey specific terminology, keywords or other contextual information to the system that might help discard false positives and make true positives more salient.

3.3 Trust

Successful adoption of automated systems requires end user trust and acceptance [5]. Our evaluation therefore focuses on the assessment of trust, acceptance of suggested extractions and the benchmarking of potential efficiency gains. There are various ways to define trust [15], and it has been shown to be related to the feeling of being in control, perceived transparency and explanation [16] to various levels, depending on the end user [17]. End users' level of trust might change over time, depending on their use of and experience with the system [16]. Our study captures trust through an explicit survey and a passive behavioural metric: user's acceptance/rejection of AI suggestions.

[1] spaCy NER demo: https://explosion.ai/demos/displacy-ent

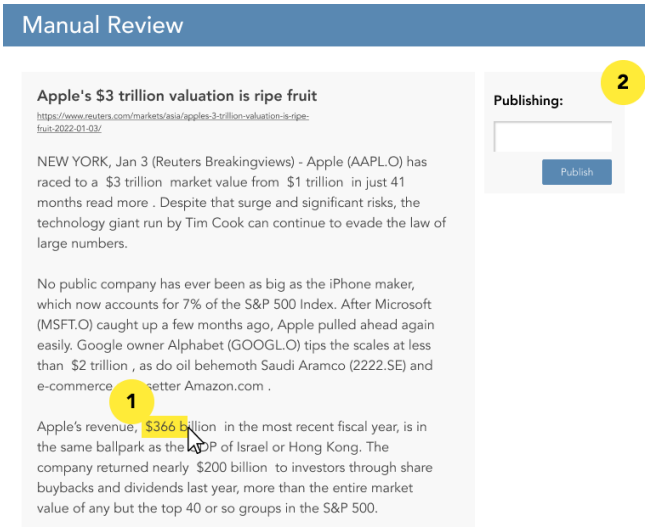

Figure 1: Manual review, (1) user has to manually find and select a "revenue value" from a full article without any ML suggestions, (2) the value selected value is copied over to the box and can be adjusted before submission

4 EXPERIMENTAL SYSTEMS

We developed an experimental system for human-assisted interactive information extraction, which enables end users to customize information extraction rules, that build on existing RegEx and NER techniques (Section 4.3). The system is subject to a patent [18]. In addition to the rule based system, we further developed three additional conditions in order to evaluate the proposed human-assisted interactive information extraction system holistically against a fully manual review baseline (Section 4.1), a basic NER extraction without an additional customization where the user is presented with multiple extractions (Section 4.2), and a black box extraction system without any explanation at all, where the user is presented with just one extraction (Section 4.4). Various conditions allowed us to compare and evaluate which system outperformed the others on different metrics (Section 5).

4.1 Manual Review

In order to benchmark the rule based review and other ML conditions that included ML assistance to a fully manual review condition, we developed an experience that allows to review a news article, manually select the text for extraction, adjust the value for submission, see Figure 1. This condition provided a baseline for efficiency and trust. No ML assistance to the user was provided in this condition.

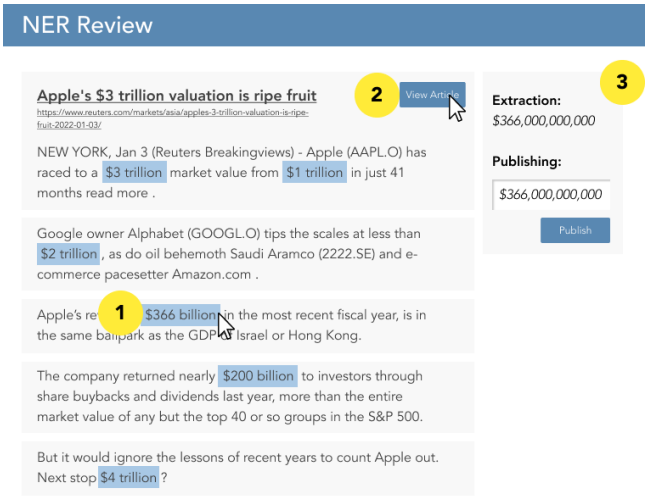

Figure 2: NER review, (1) selection from multiple suggestions based on NER extractions of "monetary values", (2) view entire article functionality (3) the value selected value is copied over to the box and can be adjusted before submission

4.2 NER Review

The Named Entity Recognition (NER) review condition extracts and displays information snippets to the user based on off-the-shelf extraction solutions for dates, monetary values and organization (via a combination of Spacy NER, datefinder[2] and custom RegEx). Instead of displaying the entire article, only sentences with relevant suggestions are displayed. Figure 2 illustrates suggestions based on the extraction of any monetary value. The user can select, adjust and submit any suggestion, or click to review the entire article on-demand, in order to double check that nothing has been missed.

Given a scenario in which a user is looking for a specific monetary value (e.g. "revenue value"), the extraction of any "money" value provides the user with high recall, but low precision, i.e. it extracts most of the relevant results, but also a large number of false positives.

4.3 Rule Based Extractor System and Review

The rule based system combines the above mentioned NER condition with an additional layer and GUI for user defined search context, e.g. keywords that should be present, or not, in target sentences, which were expressed in RegEx[3] based filters. Figure 3 illustrates the creation of a basic rule for "revenue values" that combines extraction of any "monetary value" that appears in sentences containing the keyword "revenue". Again, the extraction of all "money" values provides high recall. Adding a layer to further define the search context, should ideally increase precision as well, i.e. decrease the number of false positives. Based on the analysis of past extractions, the system suggests further improvements to the rule, e.g. filter by sentences that contain the keyword "sales", or do not contain "profit" or "shares", see Figure 3.

[2]Python Datefinder, https://pypi.org/project/datefinder/
[3]Python Regular expression operations, https://docs.python.org/3/library/re.html

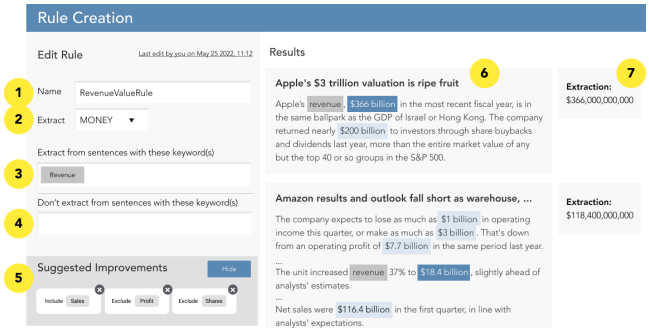

Figure 3: Rule creation, user can specify the following: (1) rule name, (2) NER extraction type, (3) keywords present in target sentences, (4) keywords absent in target sentences, (5) ML based suggestions for keywords which user can choose to accept or reject (6) example article snippet where extraction is found, (7) example extraction

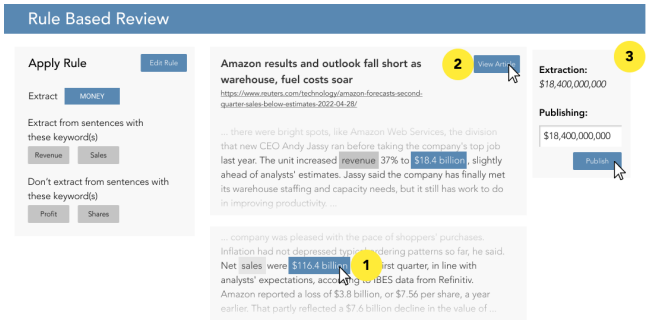

Figure 4: Rule application screen: (1) selection from a few rule based extractions, (2) possibility to view entire article, (3) adjustment of the value for submission

Figure 4 illustrates the review screen which applies a rule previously created by the user. The system only displays sentences that match with the extraction rule. It enables the user to select any of the suggested values that matches the extraction rule, review the entire article on-demand, or adjust the value. A string parsing component for the value type (e.g. "money"), not only reduces the need for reformatting to an expected publishing standard but also allows for information to be stored in a machine readable format for future analysis and enhancement.

4.4 Black Box Review

This condition allows the user to review only one suggested extraction per article representing a use case in which a larger training corpus is available and an automated black-box machine learning solution is chosen. Under the hood, the system uses a custom NER model trained on a few hundred expert-annotated examples. Most state-of-the-art high quality extraction models learn complex character and token sequence patterns that are difficult to describe to a human in a concise manner. Hence, no explanation about the inner workings of the system is given to the user, rules cannot be

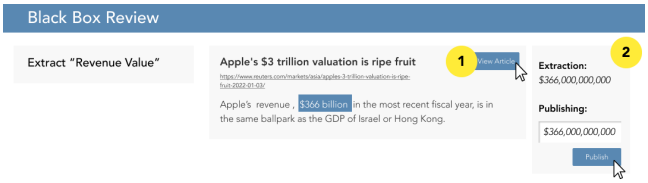

Figure 5: Black box review condition (1) view entire article functionality, (2) adjustment of the value for submission

reviewed, nor customized. The user has the ability to review the full article on-demand, in order to double check that nothing has been missed, adjust and submit the value, see Figure 5.

5 METHOD

The design of the rule based extraction application followed a user centred design approach [19]. We applied techniques such as contextual inquiry [20] and user testing of early prototypes, in order create a user experience that is usable and comprehensible to news reporters. However, the focus of this study is the evaluation of the system and a benchmark comparison between all the different conditions introduced above. More specifically their impact on end user trust and acceptance, as well as overall efficiency.

The experiment applied a within-subject design and combined observation and passive metrics with surveys and qualitative user testing. Six news reporters, who detect and extract information on a daily basis, reviewed the same set of 16 articles throughout a series of the four different conditions (Manual, NER, Rule Based, Black Box, see Section Experimental Systems) with the goal to extract the most relevant "revenue value" per article. Conditions as well as the order of the articles were randomized. Different sessions were spread across multiple days to further reduce learning effects. During each session passive metrics were captured in order to investigate time-on-task and accept/reject ratio, see Table 1. Each individual review was followed up by a questionnaire that captured a standard Single Ease Question (SEQ)[21] *"Overall, how difficult or easy was the task to complete?"* on a 5-point Likert scale, and a newly introduced a Single Trust Question (STQ) *"Overall, how trustworthy is the system?"*.

All sessions were closed with a semi-structured focus group with all participants, in which they were asked questions about the experience of each condition (e.g. *"How did you feel about the process of defining a rule"*), specific decisions (e.g. *"Why did you create rules in a certain way"*) and preference (e.g. *"How do you compare the different review conditions?"*). This method was chosen, in order to encourage exchange and discussion of the created rules and their review experience across different conditions between participants. Thematic analysis of participants' statements revealed common themes and reaction to the conditions.

6 RESULTS

Overall results show that a rule based condition significantly outperforms other conditions in terms of efficiency and end user trust, see Table 2.

Timings for these conditions were compared with a repeated measures ANOVA test and reported in mean time difference. Tukey's HSD was performed *post hoc* to a one-way ANOVA. Due to the sample sizes being large (n=73, 59, 94, 97 for Black Box/Manual/Rule based/NER), there exists good evidence that normality assumptions can be relaxed. Effect size is reported by Cohen's *D*, as it is robust across experiments. Article reviews could not be paired for the comparison with Black Box conditions. Observations for other conditions (Manual, NER, Rule Based) could be paired and therefore evaluated with a Wilcoxon paired statistic and reported in mean time difference. As Wilcoxon is a rank measure, we report biserial correlation as an indicator of effect size.

Table 1: User behaviour metrics that were tracked during the experiment

Metric	Description
Time-on-Task	Time from opening a source article to submission of extract information
No. accepted suggestions	Number of times a participant accepted extracted information without any edits
No. rejected or corrected suggestions	Number of times a participant would edit extracted information, or manually add their own

Table 2: Overview of all results, showing the overall count of all articles reviewed by all 6 participants, average time-on-task (seconds), Single Ease Question and Single Trust Question (5-point Likert scale) as well as percentage of accepted suggestions across all articles

Condition	Count	Time	SEQ	STQ	Acceptance rate
NER	97	0:26	4/5	3.1/5	93.8%
Rule	94	0:09	3.6/5	3.9/5	89.6%
Manual	59	0:14	3.6/5	3.4/5	n/a
Black Box	73	0:15	4.6/5	3.5/5	28.8%

6.1 Increased efficiency

Participants reviewed articles significantly quicker in the rule based condition as compared to any other condition, see Figure 6.

Comparing reviews in rule based and black box conditions, participants were significantly quicker (ANOVA, p-value = 0.001, Cohen's D = 0.619) with a mean time difference of -19 seconds for the rule based condition. Low efficiency in the black box condition is due to the fact that participants opened and read the full article almost 3 out of 4 times, rather than selecting a suggested extraction straight away.

Similarly, comparing reviews in rule based and NER conditions, participants were significantly quicker (Wilcoxon paired, w=300.5, p-value < 0.001) for the rule based condition with a small effect size (Biserial Correlation = 0.48) and a median time difference of -17 seconds. Low efficiency in the NER condition might relate to a larger number of false positives and general visual noise that needed review.

When comparing rule based and manual review conditions efficiency gains are the weakest with a median time difference of -4

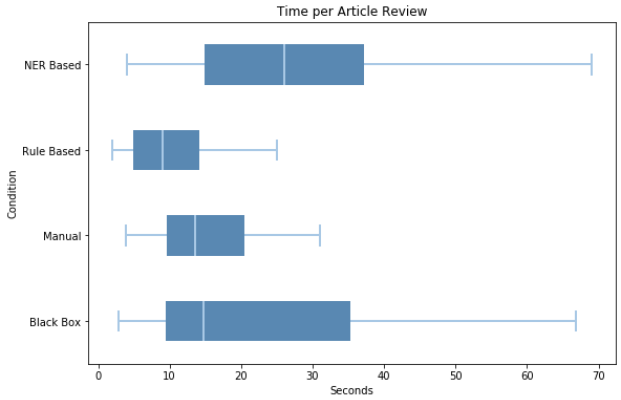

Figure 6: Results, time-on-task

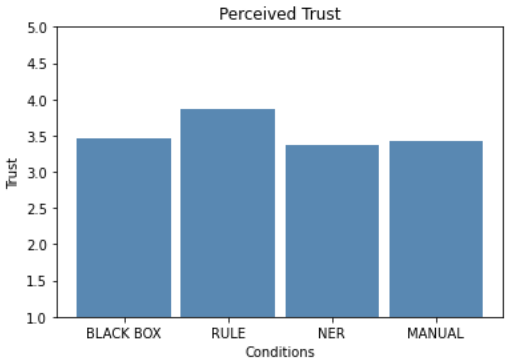

Figure 7: Results, single trust question

seconds, yet significant (Wilcoxon paired, w=45, p-value = 0.005) and a strong effect size (Biserial Correlation = 0.54).

Interestingly, a manual review tends towards being more efficient than a black box review (ANOVA, p-value = 0.054, Cohen's D = 0.34) in this experiment. Again, low efficiency for the black box condition might be due to the fact that participants opened and read the full article in this condition often as they did not trust the sole extraction that the system presented them with.

6.2 Trust, Accept/Reject ratio & Control

Participants rated their trust into the system in a rule based condition higher (3.9/5), than in black box (3.5/5), manual (3.4/5) and NER (3.1/5) conditions, see Figure 7.

Participants appeared to accept suggested extractions without opening and double checking full articles much more often in a rule based condition (89.6%), and NER condition (93.8%), as opposed to a black box condition (28.8%), see Figure 8. This can be interpreted in terms of higher level of trust into the relevance of the suggestions in rule based and NER conditions, where participants had to choose from a number of suggestions based on a simple or customized extraction rule - as opposed to the black box condition, where participants had to confirm only one suggestion without knowledge about the underlying extraction rule.

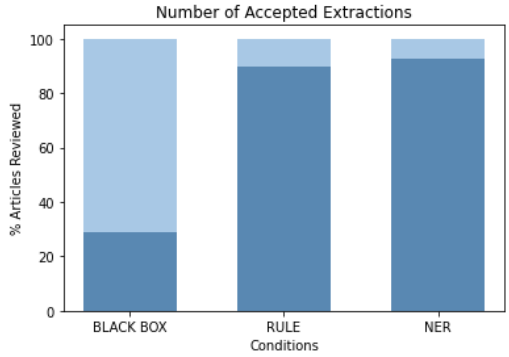

Figure 8: Results, accept/reject ratio

Table 3: Themes that emerged from collected qualitative feedback from the participants

Theme	Quotes
Control	*"I have more control over the highlights that come up – which is something I really liked about the rule based", "I want [an] amount of control over rules [...] [black box extraction] doesn't quite work. There is no way for us to set an exclude line."*
Understand	*"[the rule based condition] makes understanding and capturing [extractions] easier"*
Verify	*"[In the black box condition] you have to go back and verify", "The need to double check requires to spend a considerable amount of time"*
Context	*"I like to see incorrect and missed results. The rule based version [highlights] incorrect results. The more incorrect results you find the better you can know what is the correct result"*

Our interpretation of these results is confirmed through qualitative feedback where we heard about the need for *"being in control"*, *"understanding the system"* and *"verification of results"* and *"seeing context of suggestions"*, see Table 3.

6.3 Less ease of use

The rule based condition scored as low as manual review (SEQ = 3.6/5) in terms of ease of use, compared to high ease of use scores for NER (4/5) and black box (SEQ = 4.6/5) conditions. The creation of custom rules might have been perceived as less easy to use than a black box based review which did not require such an additional effort. However, 4 out of 6 participants stated that they preferred rule based over other conditions due to the ability for customization, "being in control" and a better understanding how the system detected values, see Table 3.

7 DISCUSSION

Our study shows that transparent and customizable rules based extraction systems can have some benefits. Participants felt comfortable to engage with the system. The experience of customizing and applying their own rule seemed to have increased participants' level

of perceived trust in the system. In comparison to other conditions participants less often reviewed and double checked the source document and therefore increasing their overall review efficiency.

We interpret that a user experience that requires the participant to make a choice from a number of suggestions, based on a more easily explainable extraction mechanism (Rule based), might be perceived as more trustworthy. Meanwhile an experience that presents only one suggestion to the user, without explaining how the suggestion was chosen (Black box) does not support such trust building interaction. While the first might evoke a feeling of perceived trust, understanding and collaboration with the system throughout the interaction, the latter might require a user to blindly trust the system and accept the suggestion. However, interaction with any kind of system, including AI systems, evolves over a period of time [22], therefore, if the users were to use all of the systems used in the experiments for longer, we may see different behaviour patterns emerging.

NER extraction underperforms compared to other conditions in terms of time-on-task. This might be due to the combination of high recall/low precision in our test scenario and hence the need to review a large number of false positives. The black box condition also underperforms in terms of time-on-task, compared to manual and rule based conditions. The time gained by extraction and presentation of one top result, was lost by participants opening and double checking the full source article more often than in other conditions. We interpret that, by not explaining the underlying mechanism, participants might have experienced less trust and might have been thrown off more easily by erroneous extractions. Our study demonstrates that for high stakes domains, where misclassified extractions might have a significant impact, it appears appropriate to keep a human closely in the loop.

Of course, custom creation of rules might have its limitations, depending on the use case. Rules that combine NER techniques and keyword search are certainly not always effective. Additional research could explore to enhance rules with syntactical rules, consideration of part of speech, fuzzy search, synonyms, filters for numerical values etc.

In addition, rules might become stale over time, if not updated regularly. However, this might be mitigated by explaining performance of individual rules, regular rule reviews, expiry dates of rules, crowd sourcing or aspects of gamification of such rules.

It also might be difficult for untrained end users to articulate rules that achieve the desired outcome. Further research could explore extending such systems with suggestions and ML based tools that could provide guard rails to end users in case their rules underperform strongly. Finally, application of such human-assisted workflows could enable capturing labelled data initially, that could be used as training data for ML based information extraction components further down the road.

REFERENCES

[1] Pamela Karr-Wisniewski and Ying Lu. When more is too much: Operationalizing technology overload and exploring its impact on knowledge worker productivity. *Comput. Hum. Behav.*, 26(5):1061–1072, September 2010.

[2] Maeda F. Hanafi, Azza Abouzied, Laura Chiticariu, and Yunyao Li. Seer: Auto-generating information extraction rules from user-specified examples. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 6672–6682, New York, NY, USA, 2017. Association for Computing Machinery.

[3] Mayank Kejriwal, Runqi Shao, and Pedro Szekely. Expert-guided entity extraction using expressive rules. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1353–1356, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. Towards explainable ai: Assessing the usefulness and impact of added explainability features in legal document summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[6] David Gunning and David Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58, Jun. 2019.

[7] Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, 38(3):50–57, Oct 2017.

[8] Artificial Intelligence at Thomson Reuters. https://www.thomsonreuters.com/en/artificial-intelligence/ai-principles.html. Accessed 8-June-2022.

[9] Artificial Intelligence at Google: Our Principles. https://ai.google/principles/. Accessed 8-June-2022.

[10] Microsoft responsible AI principles. https://www.microsoft.com/en-us/ai/our-approach. Accessed 8-June-2022.

[11] Sonit Singh. Natural Language Processing for Information Extraction. 2018.

[12] Laura Chiticariu, Yunyao Li, and Frederick R. Reiss. Rule-based information extraction is dead! long live rule-based information extraction systems! In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 827–832, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[13] Yunyao Li, Elmer Kim, Marc A. Touchette, Ramiya Venkatachalam, and Hao Wang. Vinery: A visual IDE for information extraction. *Proc. VLDB Endow.*, 8(12):1948–1951, 2015.

[14] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Jesús Bernal. A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, 26:225–238, 2012.

[15] D. Mcknight and Norman Chervany. *Trust and Distrust Definitions: One Bite at a Time*, volume 2246, pages 27–54. 01 2001.

[16] D. Holliday, S. Wilson, and S. Stumpf. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 164–168. ACM, New York, USA, March 2016.

[17] Sebastian Schnorf, Martin Ortlieb, and Nikhil Sharma. Trust, transparency & control in inferred user interest models. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, pages 2449–2454, New York, NY, USA, 2014.

[18] Hella-Franziska Hoffmann and Johannes Schleith. Systems and method for generating a structured report from unstructred data, U.S. Patent 20210232615, July 2021.

[19] Wei Xu. User centered design (vi): Human factors approaches for intelligent human-computer interaction. 2021.

[20] Hugh Beyer and Karen Holtzblatt. *Contextual Design: Defining Customer-Centered Systems.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[21] Jeff Sauro and Joseph S. Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 1599–1608, New York, NY, USA, 2009. Association for Computing Machinery.

[22] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 624–635, New York, NY, USA, 2021. Association for Computing Machinery.