

Bachelor's Thesis

Title of Bachelor's Thesis (English)	Characteristics of semi-automatic knowledge graph evaluation approaches
Title of Bachelor's Thesis (German)	Merkmale von halbautomatischen Ansätzen zur Evaluierung von Wissensgraphen
Author (last name, first name):	Liu, Oliver
Student ID number:	11704955
Degree program:	Individual Bachelor Degree Program in IBA and Chinese, BA
Examiner (degree, first name, last name):	Prof. Dr., Marta Sabou; Dipl.-Ing., Stefani Tsaneva

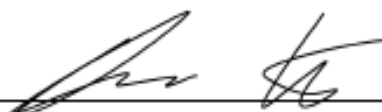
I hereby declare that:

1. I have written this Bachelor's thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced.
2. This Bachelor's Thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.
3. This Bachelor's Thesis is identical with the thesis assessed by the examiner.
4. (Only applicable if the thesis was written by more than one author): this Bachelor's thesis was written together with

The individual contributions of each writer as well as the co-written passages have been indicated.

26/10/2024

Date


Signature

Bachelor Thesis

Characteristics of semi-automatic knowledge graph evaluation approaches

Oliver Liu

Date of Birth: 15.07.1998

Student ID: 11704955

Subject Area: Knowledge Management

Studienkennzahl: J 037 561 611

Supervisor: Dipl.-Ing. Stefani Tsaneva

Date of Submission:

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

Contents

1	Introduction	7
2	Background and Related Work	9
2.1	Human in the Loop method	9
2.2	Knowledge Graph Evaluation Approaches	9
2.3	Hybrid (Human-AI) Intelligence Workflows	11
2.3.1	Hybrid intelligent systems	11
2.3.2	Moral support systems	13
3	REVIEW OF LITERATURE	16
3.1	Human Task	17
3.2	Human Role	18
3.3	Human-in-the-loop method	19
3.4	AI Task	19
3.5	AI Role	20
3.6	AI Type	21
4	DISCUSSION	23
4.1	Crowdsourcing, the dominant HITL approach	23
4.2	Cyclical human-AI interaction	23
4.3	AI to reduce human workload	24
5	CONCLUSIONS	25

List of Figures

1	Example of a triple: "Dog" linked to "Human" by "is Pet of".	7
2	How the accuracy evaluation framework functions by Qi et al. in [19]	10
3	TDP by van Zoelen et al. in [11]: AI Advisor and Human Performer. Figure in squares represents the AI and figure in circular represents the human.	12
4	TDP by van Zoelen et al. in [11]: AI Performer and Human Assistant	13
5	TDP by van Zoelen et al. in [11]: AI Performer and Human Validator	13
6	TDP by van Stijn et al. in [25]: The Human Moral Decision Maker	14
7	TDP by van Stijn et al. in [25]: Coactive Moral Decision Maker	14
8	TDP by van Stijn et al. in [25]: Suggesting Machine	14
9	Distribution of papers describing Human tasks	18
10	The Human roles amongst the 13 papers out of [21]	19
11	Distribution of papers describing crowdsourcing	19
12	Distribution of papers describing AI-Tasks	20
13	Distribution of papers describing AI-Roles	21
14	Distribution of papers describing AI-Types: the outer ring show the distribution of subsymbolic and symbolic AI; the inner ring shows the different types.	22

List of Tables

- | | | |
|---|--|----|
| 1 | Papers, which have been selected to extract information. The left column shows the reference number in this thesis. The right column shows the title of the publication. | 16 |
|---|--|----|

Abstract

This thesis explores the characteristics and methodologies of semi-automatic approaches for the evaluation of knowledge graphs (KGs). KGs play a vital role in organizing complex information, and their evaluation is crucial for ensuring accuracy and consistency, especially in fields such as artificial intelligence, decision support systems, and search engines. While fully automated KG evaluation methods exist, human intervention is often necessary to identify errors that machines might miss. This study reviews literature from 2010 to 2020, focusing on hybrid intelligence approaches that combine human expertise with automated processes, often referred to as human-in-the-loop methods. The research identifies trends, similarities, and limitations in existing semi-automatic KG evaluation methods and highlights the importance of integrating AI systems with human oversight to improve efficiency, reduce costs, and enhance accuracy. The findings provide insights into the current landscape of semi-automatic knowledge graph evaluation and offer a foundation for future research in this evolving field.

1 Introduction

A Knowledge Graph (KG) is a tool used to represent complex information in a clear and connected way. It has gained significant attention recently, especially with the growing interest in artificial intelligence (AI) and its applications in search engines and decision-making systems. A KG is typically built on top of an ontology, which serves as its foundation. An ontology is a structured schema that defines the types, properties, and interrelationships of entities at a conceptual level, organizing this information into "triples." These triples are simple statements that connect two entities (such as "Dog" and "Human") through a relationship (like "is pet of") as depicted in Figure 1. When these triples are combined in a graph format, with entities as nodes and relationships as edges, they form the structure of a KG [15].

Ontologies play a crucial role in KGs because they provide the underlying organization and semantics. They not only structure how entities and relationships are represented, but also help make data more accessible, easier to integrate, and more interpretable for tasks like data analysis and information discovery [22], [24].

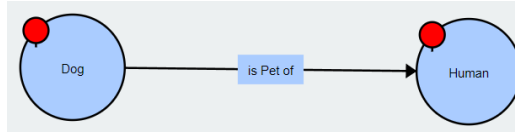


Figure 1: Example of a triple: "Dog" linked to "Human" by "is Pet of".

As these semantic resources are needed in many intelligent applications and research fields, it is of importance to ensure the correctness of these resources to avoid the failure of the applications they enable. While numerous automated methods for verifying KGs exist, certain flaws can only be detected through human intervention [12]. Those errors are typically found and corrected by domain experts. Despite the high accuracy achieved by expert evaluations, they are associated with significant costs and time requirements [13]. To address this challenge, semi-automatic approaches leveraging "Hybrid Intelligence" (HI) [16, 3] have emerged. These methods combine human intelligence with automated processes, aiming to maintain high accuracy in defect identification while minimizing evaluation time and costs. HI seeks to achieve goals that were previously unattainable by either humans or machines alone and for both to be able to improve over time.

Human Evaluation of Semantic Resources (HESR) is a critical area of study within the field of semantic technology, focusing on the involvement

of human judgment in the evaluation of semantic resources. Despite significant advancements, several challenges and gaps persist in the theoretical and practical understanding of HESR.

Sabou et al. in [21], investigated and assessed the state-of-the-art research relevant to HESR, this paper aims to address and extend the existing knowledge base and contribute to the existing literature review. Sabou et al. achieved a comprehensive understanding of the theoretical aspects of HESR, identified best practices and trends in human-involved semantic resource evaluation. However, semi-automatic approaches have not been investigated in depth according to their specific characteristics.

This thesis is a continuation of the larger literature review conducted by Sabou et al. in [21]. The goal of this thesis is to give a systematic overview of the current situation of semi-automatic approaches, relying on HI, for the evaluation of semantic resources such as ontologies and KGs. This thesis is guided by the following research question: "*What are the characteristics of current semi-automatic knowledge graph evaluation approaches?*". This question aims to address the current status quo by reviewing the existing literature on KG evaluation and identifying the different approaches currently in use. To achieve this, papers published between 2013 and 2020 were read and the relevant information was extracted. Furthermore similarities, trends and limitations are outlined. The current semi-automatic KG evaluation approaches are compared and a comprehensive overview is provided. The results of this study will be valuable to related researches in gaining a better and deeper understanding on the matter of semi-automatic evaluation of semantic resources.

The remainder of the thesis is structured as follows :

- Chapter 2: Background and Related Work - In Chapter 2 we introduce the foundations of the research domain and discuss related work.
- Chapter 3: Review of Literature - In Chapter 3 we show the extraction of the selected literature.
- Chapter 4: Discussion - Based on the extraction in the previous chapter, Chapter 4 discusses the ongoing trends.
- Chapter 5: Conclusion - In Chapter 5 we conclude the thesis.

2 Background and Related Work

We first provide an overview of KG evaluation approaches, then we describe HI workflows, focusing on team design patterns (TDP). Notable research has been done on the semi-automatic knowledge graph evaluation field and this chapter aims to give an overview.

2.1 Human in the Loop method

Human-in-the-loop (HITL) methods integrate human judgment and expertise into automated processes to improve decision-making, accuracy and adaptability in various tasks. These approaches balance the strengths of machine learning and artificial intelligence with human participation, ensuring better outcomes, especially in complex scenarios where automation alone may fall short [23]. HITL methods are widely used in fields such as data annotation, model training, and quality control, where human intervention can enhance the precision of automated systems. One prominent HITL method is crowdsourcing, which offers a crucial advantage by using the collective intelligence of a distributed workforce to perform tasks that are often simple yet essential. In knowledge graph evaluation, crowdsourcing offers a semi-automated solution by breaking down the evaluation process into manageable tasks, such as verifying entity relationships, identifying errors, classifying entities, and assessing the overall coherence and completeness of the graph [1, 26]. By involving large numbers of contributors, crowdsourcing increases accuracy and scalability, offering a cost-effective alternative to traditional methods [2, 4]. This method engages a diverse group of people, often from online communities, to complete simple tasks without requiring subject matter expertise. To distribute tasks to a broader audience, complex processes can be divided into micro-tasks, allowing for parallel task completion and more efficient solutions. Crowdsourcing platforms, such as Amazon Mechanical Turk, are commonly used for tasks like image recognition, data entry, and surveys [9, 14, 10, 8]. Expert crowdsourcing, involving individuals with domain-specific knowledge, can lead to higher-quality outcomes, more efficient solutions, and reduced performance variation, but are more costly [13].

2.2 Knowledge Graph Evaluation Approaches

There are various approaches to evaluating knowledge graphs, each offering unique methods for measuring their accuracy and performance. In this section, we aim to present exemplary methods to introduce the topic and

provide insights into effective evaluation strategies. In [19], the authors propose an interactive knowledge graph accuracy evaluation framework, which considers information extraction, as well as entity linking, at the same time, while also introducing a human-machine collaborative mechanism that leverages the data processing power of computers and the correctness verification skills of humans. Experimental validation on real and synthetic knowledge graphs underscores the potential of their approach [19]. The framework for evaluating KG accuracy, aims to minimize the total cost of sampling and human annotation, while ensuring statistical guarantees for accuracy estimates. By leveraging both the strengths of human and computer, the system proposed interleaves triple sampling and human annotation, with the machine continuously performing pre-computation during human annotations. Figure 2 shows the described accuracy evaluation framework. Inference graphs are constructed by integrating triples, entity linking results, and dependency rules, enabling accurate estimates with minimal sampling. An optimization problem is formalized to determine the optimal order of annotating triples and linkings. By utilizing feedback from annotators, the system reduces overall time costs by overlapping machine computation and human annotation time. Accuracy estimates with statistical guarantees are produced after each round of annotation[19].

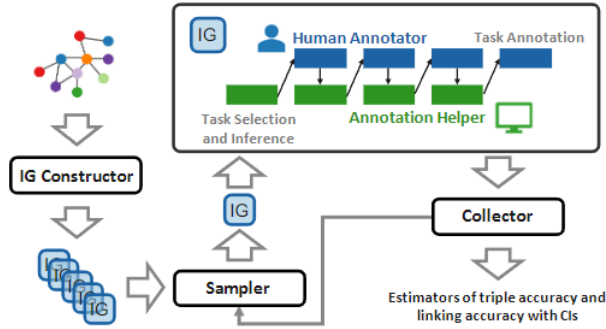


Figure 2: How the accuracy evaluation framework functions by Qi et al. in [19]

Yang et al. discusses in [27] a new approach to automated compliance checking in the construction industry. They proposed a methodology for semi-automatic construction design code knowledge graph, which includes four parts: interpretation, reconstruction, organization and implementation. The paper emphasizes the role of domain experts in dividing "original clause

text" into "semantic blocks," the fundamental units of constraint representation. This suggests a human-in-the-loop approach where experts leverage their knowledge to structure the raw code information, which likely guides the machine learning process. They concluded that a domain expert, who intervenes at the end of the semantic interpretation and knowledge reconstruction processes, is needed to review, refine, or correct the machine-generated interpretations and annotations[27].

Acosta et al. used the DBpedia dataset in [1] to conduct research on crowdsourcing mechanisms to evaluate the quality of Linked Data (LD). For their work and tests with the crowdsourcing mechanism, which specifically is intended to control human computation algorithms, by dividing complex tasks into series of easier tasks[1]. They concluded that crowdsourcing the act of detecting errors of LD in DBpedia is reasonable. While lay workers showed satisfactory precision in detecting certain issues, domain experts excelled in detecting more advanced errors, such as 'object value' or 'datatype' issues [1]. The complementary proficiencies of those two methodologies showcased is mentioned. These results could be foundational for future research in LD quality assessment using human computation.

2.3 Hybrid (Human-AI) Intelligence Workflows

The papers by Zoelen et al.[11] and van Stijn et al. [25] both focus on the concept of Team Design Patterns and their potential applications in Hybrid Intelligence (HI) systems. A TDP is the combination of text and pictorial language to describe possible solutions to recurring design problems. Both research efforts acknowledge the importance of human-AI collaboration and aim to optimize the interaction between human agents and AI systems within the TDP framework.

Both papers center around the concept of TDPs and their applicability to HI systems. They recognize the potential of TDPs in enhancing the design and development of these systems.

2.3.1 Hybrid intelligent systems

The paper by van Zoelen et al. addresses the lack of structured ways of specifying design solutions for HI systems and the absence of best practices shared across application domains. They applied this approach to three concrete HI use cases and successfully extracted team design patterns that are generalizable, providing reusable design components across various domains.

The first pattern, "AI Advisor and Human Performer" depicted in Figure 3, positions the AI as a strategic advisor, augmenting the human's decision-making process. In this pattern, the AI analyzes the task, explores various options, and presents a set of well-reasoned recommendations to the human decision-maker. The human expert then evaluates the AI's insights and ultimately makes the final call. This approach is particularly valuable when high-level human judgment and domain expertise are critical to the decision-making process, or when reducing the human's cognitive load and freeing them to focus on the most strategic aspects of the task is desired. In the sec-

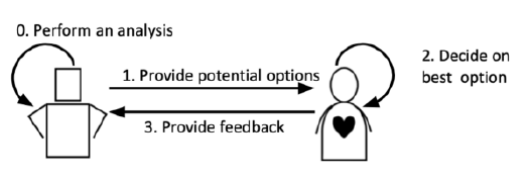


Figure 3: TDP by van Zoelen et al. in [11]: AI Advisor and Human Performer. Figure in squares represents the AI and figure in circular represents the human.

ond pattern, "AI Performer and Human Assistant" depicted in Figure 4, the AI is the primary actor, doing the bulk of the tasks. Mostly used, when the tasks are well defined, repetitive or require rapid processing of large datasets. The human expert remains on standby, ready to provide assistance when the AI encounters edge cases, needs help navigating ambiguity, or requires ethical oversight. In this pattern, the AI takes the lead in performing the task, leveraging its computational power and specialized capabilities. The human expert, however, maintains an active role, monitoring the AI's performance and stepping in when necessary. This collaborative approach allows the AI to handle the bulk of the work while the human provides guidance, oversight, and intervention when the AI faces challenges or uncertainties. The human's expertise is particularly valuable in situations where ethical considerations or complex decision-making is required, ensuring the AI's actions align with desired outcomes and organizational policies. The third pattern, "AI Performer and Human Validator" shown in Figure 5, focuses on ensuring the quality and reliability of the AI's output. The AI performs the task, but its results are then reviewed by a human expert to ensure accuracy, completeness, and alignment with desired outcomes. This is particularly relevant when accuracy is paramount, or when building trust in the AI's output is essential.

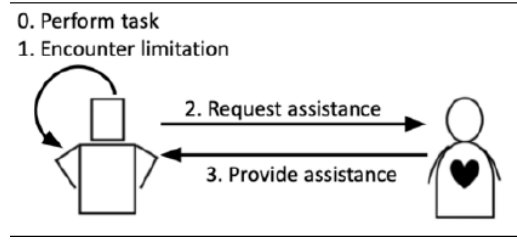


Figure 4: TDP by van Zoelen et al. in [11]: AI Performer and Human Assistant

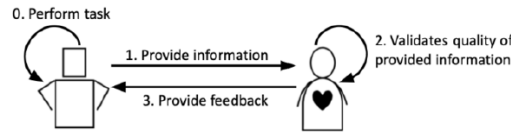


Figure 5: TDP by van Zoelen et al. in [11]: AI Performer and Human Validator

2.3.2 Moral support systems

The paper by van Stijn et al. addresses the need for ethical decision-making in medical HI systems, particularly in the context of increasing automation in the healthcare sector. The authors propose the use of TDPs to describe successful and reusable configurations of design problems where decisions have a moral component. They developed TDPs describing sets of solutions for a specific design problem in a medical HI system. A survey was created to assess the usability of the patterns with regards to their understand-ability, effectiveness, and generalize-ability.

The first pattern, "Human Moral Decision-Maker" shown in Figure 6, has the AI positioned as a consultant. It analyzes the situation and offers recommendations, the human maintains complete control and decides on the final action. If the human thinks that the suggestion is wrong, the machine will stop its task and the model can be changed according to the humans decision.

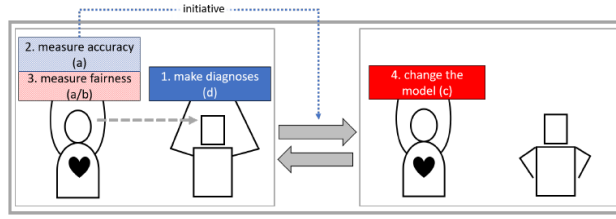


Figure 6: TDP by van Stijn et al. in [25]: The Human Moral Decision Maker

In the second pattern, "Coactive Moral Decision Maker" depicted in 7, the AI takes the lead in executing the task, while the human provides oversight and assistance when needed. The AI is programmed to recognize its limitations and will flag any uncertainties or complexities for human intervention. This pattern excels at automating routine tasks, freeing human experts for more nuanced challenges.

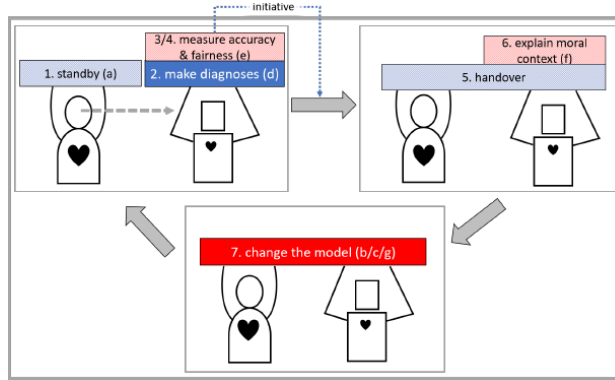


Figure 7: TDP by van Stijn et al. in [25]: Coactive Moral Decision Maker

The third pattern, "The Suggesting Machine" in Figure 8, is designed to mitigate bias in HI systems. The machine focuses on suggesting TDP that can help address potential biases. The machine decides whether a change is necessary and suggests the human different options to cancel the bias. The human reviews the suggestions and picks the method to change the model.

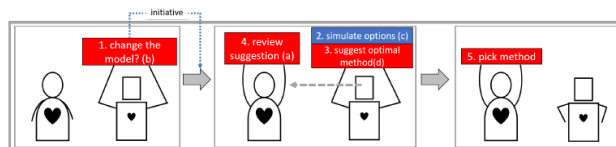


Figure 8: TDP by van Stijn et al. in [25]: Suggesting Machine

Van Zoelen and van Stijn both share the conclusion, that TDPs are useful methods to describe solutions to design problems within HI systems that involve moral decision-making. They emphasize the effectiveness of TDPs in communication amongst multidisciplinary teams and identified TDPs as a useful instrument for human-AI collaboration with a strong emphasis on the adaptability and re-usability of design patterns. The papers differ in their scope and approach. Zoelen et al. concentrate on developing and evaluating a methodology for creating TDPs; in contrast, van Stijn et al. apply TDPs to address moral decision-making in a specific case study involving bias mitigation in a healthcare system, employing the Scenario-Based Elicitation methodology and incorporating value-sensitive design principles. Consequently, while both papers explore TDPs within the context of HI, Zoelen et al. provide a broader framework for TDP development, while van Stijn et al. offer a more focused application in the realm of ethical decision-making.

3 REVIEW OF LITERATURE

Out of the bigger previous literature review [21], 13/100 papers have been identified as describing a semi-automatic approach and have been thus selected for an in-depth literature review in this thesis. The selected papers are listed in Table 1.

Table 1: Papers, which have been selected to extract information. The left column shows the reference number in this thesis. The right column shows the title of the publication.

Reference Number	Publication Title
[4]	Large-scale linked data integration using probabilistic reasoning and crowdsourcing
[20]	The BBC World Service Archive Prototype
[7]	Combining information extraction and human computing for crowd-sourced knowledge acquisition
[17]	Exploiting users' feedbacks: Towards a task-based evaluation of application ontologies throughout their lifecycle
[2]	KATARA: A Data Cleaning System Powered by Knowledge Bases and Crowdsourcing
[8]	Refining Automatically Extracted Knowledge Bases Using Crowdsourcing
[10]	Use of Ontology Structure and Bayesian Models to Aid the Crowdsourcing of ICD-11 Sanctioning Rules
[14]	Kgeval: Accuracy estimation of automatically constructed knowledge graphs
[1]	Detecting Linked Data quality issues via crowdsourcing: A DBpedia study
[9]	OC-2-KB: integrating crowdsourcing into an obesity and cancer knowledge base curation system
[5]	Efficient Knowledge Graph Accuracy Evaluation
[18]	You are Missing a Concept! Enhancing Ontology-Based Data Access with Evolving Ontologies
[19]	Evaluating Knowledge Graph Accuracy Powered by Optimized Human-machine Collaboration

To find similarities and differences amongst the papers, 6 characteristics of the semi-automatic eval. approaches have been identified and categorise.

Concretely, details on the following characteristics were extracted:

- Human Task - What the human task was and how the task helped to achieve each papers' goals.
- Human Role - The role in the semi-automatic workflow.
- Human-in-the-loop method - Whether or not human labor (crowd-sourcing) was used to achieve the results.
- AI Task - How the AI contributed to the end results.
- AI Role - Here, the tasks are categorized into concepts, which describe the task done by the AI.
- AI Type - Throughout the papers, the AI has different tasks, which are done by specific AI types. Those are separated and will be classified.

The number of extracted roles and tasks will be more than the amount of articles due to the fact, that the human or the AI often have more than one task or role.

3.1 Human Task

The role of humans varied across the studies: while some gave the human just one specific task, others had bigger workloads, which were separated into more smaller ones. This is the reason why the total amount of tasks exceeds the number of papers. Figure 9 shows the distribution of the 3 tasks "Data Evaluation", "True/false labeling" and "Improvement".

Out of the 13 papers, 10 papers had the human do true/false labeling, which includes the validation of semantic resources such as triples and beliefs. In [20, 7, 2, 8, 14, 1, 9, 19], the human checks the correctness of the extractions made by the AI, encompassing tasks such as confirming the accuracy of proposed data repairs [2], refining KB entries [19] and verifying triples and entity linkages by reviewing source texts, like extracted information or suggested changes - ensuring the integrity of the data and guiding subsequent actions [2, 9, 8, 19]. In contrast, [10, 5] validate the correctness of decisions made by other humans, e.g. through crowdworkers (CW).

Data Evaluation is performed in 5 papers. It is used for when the human generates and processes data [4, 5, 17, 18, 19], does instance matching [4] or annotations of extracted information.

The task improvement, seen in 7 papers, involves assessing the effectiveness of changes, enhancing the completeness and quality of the KG. In

semi-automatic KG evaluation, the system is partly automated, yet human intervention is required at some stages [20, 7, 17, 2, 1, 9].

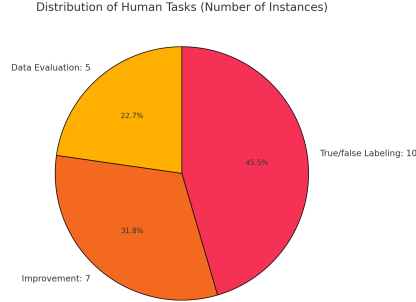


Figure 9: Distribution of papers describing Human tasks

3.2 Human Role

The human’s role is multifaceted - they serve as domain experts, providing valuable insights and knowledge to assess the validity and relevance of captured information[6]. Humans furthermore can manually curate and correct errors in the KG, which ensures overall quality and reliability. For the information extraction, the role of the human has been categorized in "Validation Performer", "Validation Reviewer" and "Curator", as seen in Figure 10. The validation performer (seen in 8 publications) checks automatically or manually extracted triples by themselves[4, 20, 17, 10, 14, 1, 5, 18]. Noted in 8 papers, the validation reviewer focuses on verifying outputs, which were generated by AI systems, ensuring alignment with human judgement and domain expertise [7, 2, 8, 9, 19]. In [18, 4, 20], the human was also tasked with performing, as well as reviewing, hence the iteration. By doing validation work errors can be detected and feedback to improve the AI’s performance can be given. The role of a curator was present in 4 publications. Here, the human refines and builds models, adds annotations, which will be used to train the AI[20, 1, 9, 19].

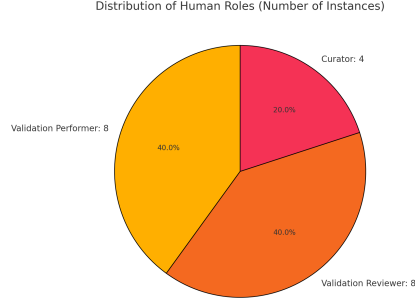


Figure 10: The Human roles amongst the 13 papers out of [21]

3.3 Human-in-the-loop method

It is notable that 10 out of 13 papers (Figure11) were using the Human-in-the-loop (HITL) method. With the HITL-method, humans remain actively involved in the tasks, done by AI-systems, including decision-making processes or guiding the AI-systems.

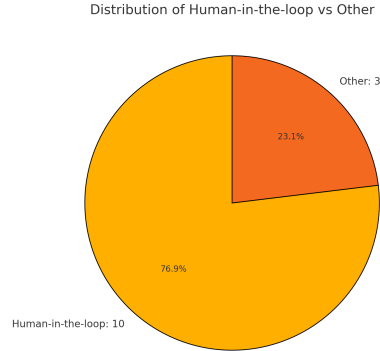


Figure 11: Distribution of papers describing crowdsourcing

3.4 AI Task

AI systems perform a wide array of tasks across domains such as information extraction, ontology management, and KB construction, enabling automation, accuracy, and curation of information from structured and unstructured data. The tasks have been grouped into "Extraction", "Entity Matching", "Rule-based Anomaly detection" and "Optimization" as seen in Table 12. Extraction had 8 instances [4, 20, 7, 2, 8, 1, 9, 14] due to the broad definition base. The term was used to describe KG construction, information

extraction and entity extraction. Validation work and inferring correctness of additional beliefs. The task entity matching, seen in 5 papers, refers to the usage of AI when identifying and linking records from different data sources that refer to the same real-world entity[4, 7, 9, 18, 19]. Rule based anomaly detection, used in 7 papers, refers to using predefined rules and thresholds to identify unusual or abnormal behavior or patterns in the system’s operations. It is used to describe tasks like consistency checking, reasoning and selection of annotations [19, 18, 5, 1, 14, 10, 17]. The optimization task is used in 3 and describes pattern discovery, cost optimization and annotation order optimization[2, 10, 5].

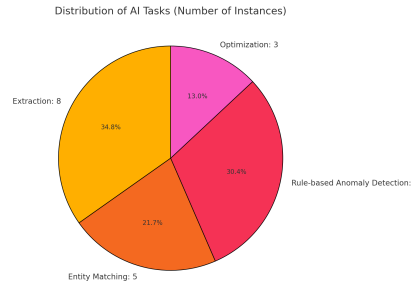


Figure 12: Distribution of papers describing AI-Tasks

3.5 AI Role

The role of the AI have been defined into distinct functions like "Extractor", "Validation Assistant", "Validation Performer" and "Optimizer", visualized in Figure 13. For the extractor every instance has been selected, where the AI extracts knowledge from structured or unstructured information and assists in annotations[4, 20, 2, 8, 1, 18, 7, 9] for a total of 8 papers. The validation assistant role, seen in 6 publications, describes AIs, which supports human experts or crowdworkers during validation of automatically evaluated results. It evaluates the ontology for consistency after changes suggested by the user or the AI [17, 10, 14, 5, 18, 19]. In contrast the validation performer, noted in 5 papers, does the validation tasks [4, 8, 9, 2, 7], where the human takes the reviewer role, as described in 3.2. The last AI role, the optimizer, focuses on refining the efficiency and effectiveness of the overall model. This role appears in 6 out of 13 papers. Under this falls semantic pattern discoverer and suggester of possible procedures [2, 5], models which improve the accuracy of rulesets [10, 14, 18] and creation of micro tasks to support the human-in-the-loop approach [1].

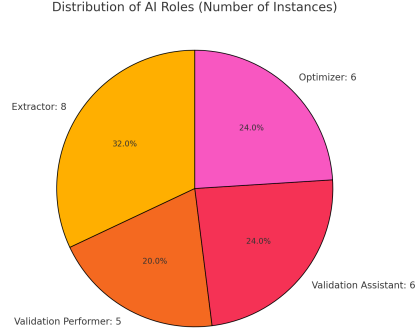


Figure 13: Distribution of papers describing AI-Roles

3.6 AI Type

Throughout the papers, the used AI did not differ a lot. The types can generally be classified into symbolic and subsymbolic AI as pictured in figure 14. The subsymbolic AI relies on data driven approaches, often focusing on patterns and statistical correlations rather than explicitly defined rules. Under this category fall machine learning (ML), natural language processing (NLP), probabilistic/statistical AI. ML is used for automated extraction of knowledge and fact correctness[8], automated tagging and integrate confirmed knowledge into the databases [4, 20], generating micro tasks and integrating data into automated process [1] and detecting relations and classification tasks within a KB [9, 19]. NLP is used for entity recognition and linking processes [4] and extracting information/relationships from literature [7, 9]. Probabilistic/statistical AI utilize probability models or statistical models to evaluate and integrate data [4, 5]. The symbolic AI used in [18, 17, 2] acts as a logical reasoner and performs validation checks.

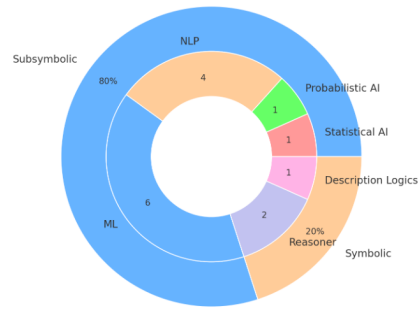


Figure 14: Distribution of papers describing AI-Types: the outer ring show the distribution of subsymbolic and symbolic AI; the inner ring shows the different types.

4 DISCUSSION

The semi-automatic approach to evaluating and refining KGs relies on a synergy between AI and human input, often in the form of crowdsourcing. This method is able to use the efficiency and scalability of AI to handle large datasets, while relying on humans for tasks requiring judgment and expertise beyond the AI’s capabilities. As demonstrated in the 13 studies in Table 1, a consistent pattern of combining human and machine can be seen to improve data quality and achieve cost-efficient scalability.

4.1 Crowdsourcing, the dominant HITL approach

One noticeable and recurring trend in the literature is the reliance on crowdsourcing for HITL evaluation tasks. Crowdsourcing can be applied at scale on various tasks such as fact verification, error detection, and refinement of AI-generated outputs to a large number of non-expert human workers. Human workers play a crucial role in validating and refining the output of AI systems, especially at tasks, where contextual understanding or subjective judgment is needed. The preference for using crowdsourcing allows systems to manage large datasets without overburdening a small group of expert annotators. This makes crowdsourcing especially appealing for semi-automatic approaches, where the goal is often to process data at scale, leveraging humans to handle cases of ambiguity or uncertainty that AI cannot resolve itself. Furthermore, as AI systems become more and more advanced, the crowdsourcing tasks become increasingly focused on higher-level judgment and curation of the AI rather than simple data validation. This was noticeable as the papers published later were using the crowdworkers (CW) output to train the AI.

4.2 Cyclical human-AI interaction

Another common pattern was the cyclical nature between AI and human evaluators. Many studies started with AI processing or extracting information, followed by human validation and then returns back to the AI for refinement based on the input given by the human work. This allows AI systems to continuously improve the accuracy based on learning from human input, which in return enhances the KGs quality.

4.3 AI to reduce human workload

A key advantage of semi-automatic approaches is the way AI systems reduce the cognitive load on human workers, while also reducing costs. AI allows human contributors to focus on tasks that require more nuanced judgment, such as validating complex relationships or identifying subtle inconsistencies in data, by automating the extraction and processing of large datasets. Furthermore, recent studies have highlighted that task flows beginning with AI-led error identification, followed by human review and ending with AI confirmation or updating, not only reduce the burden on CWs but also minimize redundancy and increase overall efficiency. In contrast, workflows that involve CWs in early stages, without prior AI pre-processing, have been shown to increase both costs and time without significantly improving outcomes. This refined task design showcases the growing emphasis on intelligent collaboration between AI and human workers to maximize efficiency and reduce manual curation, reflecting an important shift toward sustainable, scalable HITL systems.

5 CONCLUSIONS

This thesis addressed the problem of evaluating knowledge graphs, which are crucial for organizing complex information, but often require human intervention for accurate validations. The primary research question was: "What are the characteristics of current semi-automatic knowledge graph evaluation approaches?". The research aimed to systematically review existing literature on hybrid intelligence systems and their application in knowledge graph evaluation, focusing on integrating human expertise with automated processes. The methodology involved a comprehensive literature review of studies published between 2010 and 2020, extracting data on semi-automatic evaluation approaches and analyzing the role of both AI systems and human-in-the-loop methods. Key tasks, roles, and interactions in human-AI workflows were categorized and compared to identify trends and limitations. In conclusion, this thesis provides an overview of semi-automatic methods, which combine the strengths of automated AI systems with human expertise. Particularly through HITL approaches, the accuracy and efficiency of KG evaluations can be significantly improved. The cyclical interaction between humans and AI, where machines handle large-scale data processing and humans refine ambiguous or complex cases, ensures a cost-effective and scalable solution. This hybrid approach not only reduces the time and cost associated with manual evaluations but also enhances the quality of knowledge graphs by leveraging both computational power and human judgment. The reviewed papers exemplify the trend towards more collaboration and integration between AI and humans. Future research should explore more refined frameworks for optimizing this collaboration, focusing on minimizing redundancy and further improving the scalability of knowledge graph evaluation systems.

References

- [1] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Fabian Flöck, and Jens Lehmann. Detecting linked data quality issues via crowdsourcing: A dbpedia study. *Semantic web*, 9(3):303–335, 2018.
- [2] Xu Chu, John Morcos, Ihab F Ilyas, Mourad Ouzzani, Paolo Papotti, Nan Tang, and Yin Ye. Katara: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1247–1261, 2015.
- [3] Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354*, 2021.
- [4] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal*, 22(5):665–687, 2013.
- [5] Junyang Gao, Xian Li, Yifan Ethan Xu, Bunyamin Sisman, Xin Luna Dong, and Jun Yang. Efficient knowledge graph accuracy evaluation. *arXiv preprint arXiv:1907.09657*, 2019.
- [6] Xiou Ge, Yun Cheng Wang, Bin Wang, C-C Jay Kuo, et al. Knowledge graph embedding: An overview. *APSIPA Transactions on Signal and Information Processing*, 13(1), 2024.
- [7] Sarath Kumar Kondreddi, Peter Triantafillou, and Gerhard Weikum. Combining information extraction and human computing for crowd-sourced knowledge acquisition. In *2014 IEEE 30th International Conference on Data Engineering*, pages 988–999. IEEE, 2014.
- [8] Chunhua Li, Pengpeng Zhao, Victor S Sheng, Xuefeng Xian, Jian Wu, and Zhiming Cui. Refining automatically extracted knowledge bases using crowdsourcing. *Computational Intelligence and Neuroscience*, 2017(1):4092135, 2017.
- [9] Juan Antonio Lossio-Ventura, William Hogan, François Modave, Yi Guo, Zhe He, Xi Yang, Hansi Zhang, and Jiang Bian. Oc-2-kb: integrating crowdsourcing into an obesity and cancer knowledge base curation system. *BMC medical informatics and decision making*, 18:115–127, 2018.

- [10] Yun Lou, Samson W Tu, Csongor Nyulas, Tania Tudorache, Robert JG Chalmers, and Mark A Musen. Use of ontology structure and bayesian models to aid the crowdsourcing of icd-11 sanctioning rules. *Journal of Biomedical Informatics*, 68:20–34, 2017.
- [11] P Lukowicz et al. Developing team design patterns for hybrid intelligence systems. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, volume 368, page 3. IOS Press, 2023.
- [12] André Melo and Heiko Paulheim. An approach to correction of erroneous links in knowledge graphs. In *CEUR Workshop Proceedings*, volume 2065, pages 54–57. RWTH Aachen, 2017.
- [13] Jonathan M Mortensen. Crowdsourcing ontology verification. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 448–455. Springer, 2013.
- [14] Prakhar Ojha and Partha Talukdar. Kgeval: Accuracy estimation of automatically constructed knowledge graphs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1750, 2017.
- [15] Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *Artificial Intelligence Review*, 56(11):13071–13102, 2023.
- [16] Niccolo Pescetelli. A brief taxonomy of hybrid intelligence. *Forecasting*, 3(3):633–643, 2021.
- [17] Perrine Pittet and Jérôme Barthélémy. Exploiting users feedbacks-towards a task-based evaluation of application ontologies throughout their lifecycle. In *International conference on knowledge engineering and ontology development*, volume 2, pages 263–268. SCITEPRESS, 2015.
- [18] André Pomp, Johannes Lipp, and Tobias Meisen. You are missing a concept! enhancing ontology-based data access with evolving ontologies. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 98–105. IEEE, 2019.
- [19] Yifan Qi, Weiguo Zheng, Liang Hong, and Lei Zou. Evaluating knowledge graph accuracy powered by optimized human-machine collaboration. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1368–1378, 2022.

- [20] Yves Raimond, Tristan Ferne, Michael Smethurst, and Gareth Adams. The bbc world service archive prototype. *Journal of web semantics*, 27:2–9, 2014.
- [21] M Sabou, M Fernandez, M Poveda-Villalón, MC Suárez-Figueroa, and S Tsaneva. Human-centric evaluation of semantic resources: a systematic mapping study. *preparation for ACM Comp. Surveys*, 2024.
- [22] Angelo A Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Aliaksandr Birukou, Francesco Osborne, and Enrico Motta. The computer science ontology: A comprehensive automatically-generated taxonomy of research areas. *Data Intelligence*, 2(3):379–416, 2020.
- [23] Luciano Serafini, Raul Barbosa, Jasmin Grosinger, Luca Iocchi, Christian Napoli, Salvatore Rinzivillo, Jacques Robin, Alessandro Saffiotti, Teresa Scantamburlo, Peter Schüller, et al. On some foundational aspects of human-centered artificial intelligence. *arXiv preprint arXiv:2112.14480*, 2021.
- [24] Marta Contreiras Silva, Patrícia Eugénio, Daniel Faria, and Catia Pesquita. Ontologies and knowledge graphs in oncology research. *Cancers*, 14(8):1906, 2022.
- [25] Jip J van Stijn, Mark A Neerincx, Annette ten Teije, and Steven Vethman. Team design patterns for moral decisions in hybrid intelligent systems: A case study of bias mitigation. In *CEUR Workshop Proceedings*, volume 2846. CEUR-WS, 2021.
- [26] Yufei Xie, Xiang Liu, and Qizhong Yuan. Research on college students’ innovation and entrepreneurship education from the perspective of artificial intelligence knowledge-based crowdsourcing. *arXiv preprint arXiv:2212.05906*, 2022.
- [27] Mingsong Yang, Qin Zhao, Lei Zhu, Haining Meng, Kehai Chen, Zongjian Li, and Xinhong Hei. Semi-automatic representation of design code based on knowledge graph for automated compliance checking. *Computers in Industry*, 150:103945, 2023.